Available online at www.sciencedirect.com

ScienceDirect

Corrosion Science 49 (2007) 4415–4442

CORROSION SCIENCE

www.elsevier.com/locate/corsci

Evaluation of image segmentation approaches for non-destructive detection and quantification of corrosion damage on stonework

P. Kapsalas [a], M. Zervakis [a,*], P. Maravelaki-Kalaitzaki [b]

[a] *Technical University of Crete, Department of Electronics & Computer Engineering, Chania 73100, Greece*
[b] *Ministry of Culture, 25th Ephorate of Prehistoric & Classical Antiquities, Chania 73100, Greece*

Received 31 July 2006; accepted 23 March 2007
Available online 20 August 2007

Abstract

This paper approaches the non-destructive analysis of corrosion damage by testing and evaluating several image segmentation schemes for the detection of decay areas. The application test bed for algorithmic evaluation considers stonework surfaces for corrosion damage. Each of the detection approaches handles in a different way the background inhomogeneities. A semi-automated framework for validating the algorithms' performance is introduced. This framework guarantees reliable and objective estimation of algorithms' response, while also enabling informed experimental feedback for the design of improved segmentation algorithms. Further to elaborating on the robust points of each segmentation approach, this work also studies the corrosion mechanisms. The latter process involves investigation of the degradation state as reflected by the size of the decay areas and their darkness. The derived assessments closely converge to assessments based on chemical analyses, performed on the same stone surfaces.

Keywords: B. Image segmentation algorithms; B. Statistical tests; B. Performance evaluation; C. Stone decay; C. Black crust

[*] Corresponding author. Tel.: +30 2821037206.
E-mail address: michalis@danai.systems.tuc.gr (M. Zervakis).

0010-938X/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.corsci.2007.03.049

1. Introduction

The degradation phenomena encountered on stonework form an aspect of high importance nowadays. Several investigations were carried out with the aim of studying the factors, extent and phenomenology of stone decay [1]. In a polluted environment, the most frequently observed decay phenomenon on stone surfaces is the formation of black crust [2]. Black crust is a coating with color ranging from reddish-brown to brown-black depending on the lithotype and the way of exposure to environmental agents. Analyses of black crusts reveal gypsum, residual calcite, silicates, potassium nitrate, iron oxides, mica flakes, quartz and numerous organic constituents in low concentration [3]. The thickness of black crust ranges from 100 μm up to several mm; successive sub-layers can be discriminated into the crust matrix, differing in texture and chemical composition. Small white particles, of gypsum crystals and re-crystallized $CaCO_3$, as well as unevenly distributed black carbonaceous particles, responsible for the coloration, were identified on the body of black crust.

Black crust not only alters the aesthetic view of stonework, but also leads to further corrosion due to catalyzing actions. Thus, the employment of chemical cleaning methods becomes essential both for conserving the artwork and preventing further degradation phenomena. A critical issue in the corrosion-damage estimation is the objective and accurate segmentation of degraded areas. Several methodologies of corrosion-damage estimation have been developed in recent years. The primary procedure is based on the ablation of the examined stone material and a subsequent chemical analysis to reveal the extent and types of degradation. However, this process is destructive to the material and increases the stone exposure to weathering conditions. The implementation of non-destructive techniques capable of providing reliable results concerning both the topology and extent of degradation is currently an aspect under investigation.

Most of the research efforts relative extracting information from corroded areas on artworks, are focused towards assessments of degradation effects on old paintings and less on detecting corrosion damage on stonework. This is mainly due to the diversity of the encountered corrosion effects (flaws, material loss, discoloration, black crusts, etc.) and the lithotype variations. A non-destructive technique introduced by Moropoulou and Avdelidis [4], assessed the corroded untreated, as well as the treated areas on stonework via an infrared thermography screening system. An automated approach implemented by Lebrun et al. [5], quantified corrosion damage through color alteration by computing the Euclidean distance in a (pseudo)-$L^*a^*b^*$ color space. Many investigations for accurate geometric analysis of the decayed material surfaces have also been proposed using topographical acquisition data. Gelli and Virulano [6] employed an automated approach known as "Shape from Shading Method" to perform reconstruction of degraded stone surfaces. Furthermore, methods for characterizing the stone structure and detecting regions of material loss were developed in the study of Moltedo et al. [7], while Boukouvalas et al. [8] introduced computer vision techniques for the detection and classification of mineral veins on ceramic tiles. Pappas and Pitas [9] employed image-processing approaches with the objective of diagnosing corrosion defects on old paintings and performing reconstruction of the digital image on the locations of degraded areas. The way that corrosion damage affects the structural integrity of aerospace materials has also been investigated. Many computer vision techniques have been developed to non-destructively detect decay areas on metals and to classify corrosion patterns according to their origin and type. An early attempt of segmenting degraded areas on metals was performed in [10], where the decay effects are inspected by eddy currents and infrared thermography. The information gathered is subsequently fused with the use of statistical and/or probabilistic algorithms. Another approach aiming at recognizing corroded areas on aerospace materials and classifying them according to their type was introduced by Choi and Kim [11]. A similar study reported in [12] is focused towards recognizing the various defects encountered on a cold mill strip.

In this paper, we focus on the performance evaluation and the potential of several image segmentation algorithms in correctly detecting and localizing decay effects. Previous performance evaluation of segmentation algorithms has been largely based on manual tuning of algorithms' parameters. We present an automated approach for extracting the Ground Truth Matrix (GT) of decay areas segmented by different algorithms and for evaluating their performance. The use of multiple sets of images provides the test bed for detecting significant performance differences among algorithms. Such an approach makes it possible to objectively and reliably compare the performance of segmentation processes, while allows for informed experimental feedback towards the design of improved algorithmic schemes.

Through the algorithms' performance evaluation the role of experts is critical. The experts pose the criteria for defining the algorithmic performance and comparing algorithms in terms of their efficiency to provide reliable results of the decay topology and extent. These criteria essentially determine the features of an appropriate segmentation approach. Usually the expert's opinion is absorbed in the process of Ground Truth Matrix estimation. In this work, we consider the potential of a detector to segment all susceptible areas (even those that do not correspond to decay effects) as indicative of its efficiency. Such a response is considered preferable by the experts. Besides the comparison of several algorithmic approaches, in this paper we investigate how exposure or even cleaning conditions are reflected in the size and the relative intensities of the corroded areas (over the background). This aspect is approached by using statistical tests to assess the significance of discrepancies observed in the decay characteristics of the examined structures. These tests also contribute in evaluating the efficiency of chemical cleaning and understanding the procedures of decay evolution. The testing framework includes image data sets of degraded stone surfaces screened by the fiber optics microscope (FOM) and digital camera. The specific images used to assess the potential of the algorithmic schemes are selected to reflect the decay phenomenology encountered on stonework.

Structuring the paper, we present in Section 2 the experimental setup, as well as the requirements considered for effective segmentation. Section 3 presents the algorithms tested and the details of the procedure developed for GT extraction. Moreover, it discusses the statistical measures employed for evaluating the algorithmic results. The results of the algorithmic testing along with the details of structural and cleaning effects on corroded surfaces are presented in Section 4. Finally, Section 5 summarizes this work and discusses directions of further required research efforts.

2. Experimental setup

2.1. Problem specification

The studied images represent degraded stone regions monitored via a FOM system and a digital camera. The FOM images depict sheltered and unsheltered areas obtained from the columns of the National Archaeological Museum (Athens), while the digital camera images represent a stone specimen depicting adjacent cleaned and uncleaned strips. The FOM images are further subdivided into reedings and flutings, in order to study the different degradation and structural effects encountered on surfaces of different exposure to weathering conditions. Thus, reedings represent areas more exposed to the rain and winds' action and consequently the black crusts occurring on these areas tend to be thinner than the corresponding crusts encountered on the adjacent flutings surfaces. Moreover, reedings present flaws and granular texture due to the removal of stone grains through the water's fluency. On the other hand, unsheltered surfaces tend to develop more lamellar texture and crusts thinner in thickness. The latter observation can be explained by taking into account the water activity that results in removing the deposited materials. Furthermore, the discoloration of the unsheltered surfaces and the formation of reddish-brown or brown-black strains are attributed to effects of dissolution of the substrate due to the water's action. We assess the severity of degradation in terms of the size of the detected decay areas and the alteration of the relative (over the background) intensities on areas of corrosion damage.

As it was previously discussed, this paper also aims at examining the effects of cleaning interventions. The applied cleaning treatments on the FOM images include: (a) an ion-exchange resin paste with deionized water (DS), (b) a biological paste (BP) comprising 1000 ml of deionized water, 50 g of $(NH_2)_2CO$, 20 ml of $(CH_2OH)_2CHOH$ and approximately 800 g of sepiolite, and (c) a wet microblasting method (WMB) springing spherical particles of calcium carbonate with diameter lower than 80 μm at a maximum function pressure of 0.5 bar; the proportion of water and spherical particles of calcium carbonate in the device's commixture barrel was 3:1. In order to assess the cleaning performance, chemical investigations with the aid of destructive techniques, such as Fourier transform infrared spectroscopy (FTIR), scanning electron microscopy with energy dispersive X-ray analysis (SEM-EDS) and X-ray diffraction analysis (XRD) were also performed on the cleaned surfaces. The results of the chemical analyses are subsequently used as input to the statistical tests in order to estimate the effectiveness of the cleaning methods in removing decay. The methods presented have been tested in 12 images of untreated surfaces, seven images depicting stone regions cleaned by the DS method, five images representing stone surfaces treated by the BP and seven WMB cleaned images.

Aiming at assessing the effectiveness of the segmentation algorithms in detecting various types of corrosion defects we selected three representative images (two FOM images and one digital camera image), where we also extracted the GTs. The images were selected with the aid of the experts, to reflect the deterioration encountered in a variety of environmental conditions. The chosen images correspond to sheltered and unsheltered untreated flutings representing structures of different texture and decay extent. Furthermore, the spatial arrangement of degradation particles was also considered for the selection of the two FOM images.

The digital camera system is also recruited as to investigate the algorithms' potential in accurately segmenting decay areas on images at low feature resolution. The image screened via this modality corresponds to a stone surface where adjacent strips of laser cleaned and uncleaned areas occur. The cleaning process was conducted by a Nd:YAG laser system used to partially remove the crust [13]. The energy fluency of the Nd:YAG laser was fixed at 6.3 J/cm^2. Throughout the cleaning process, some parameters of the laser pulses were modified, resulting in the removal of crust layers differing in thickness. Each cleaned strip was obtained by increasing the number of laser pulses per spot from one up to six; a 40% area overlap was recorded between adjacent spots.

2.2. Principles and requirements for segmentation

The development of algorithmic approaches that can accurately detect the location and structure of corroded areas aids the reliable assessment of decay phenomena. The presence of noise in the images, as well as the inhomogeneity of the stone structure, leads to the induction of false positive and false negative segmented spots. The presence of such spots may alter the estimation of decay effects, so that the restriction (or even elimination) of false segments is of high importance.

In order to design a detector that performs accurate localization of decay areas, the peculiarities of the problem should be clearly identified as follows:

- The objects of interest are very small; they are visible as dark particles in the image.
- They often appear in an inhomogeneous background that reflects the structure of the marble surface. The background structure may be darker in some parts of the image than the decay particles on other parts. For this reason, a simple thresholding scheme cannot be used for segmentation. The employed detector should take under consideration the local characteristics of the image.
- Sub-areas that depict a non-uniformity of the underlying texture are more susceptible to be decayed.
- Another problem is the usual low contrast between the objects of interest and the background. This contrast is sometimes comparable to noise contrast caused by the inhomogeneity of the stone structure. Due to the random growth of decay patterns, there is no lower bound to their contrast over the background. Obviously, the aim should be to design the segmentation process as sensitive as possible to the systematic variations caused by deterioration patterns, while suppressing those random variations caused by noise. This means that segmentation has to take dynamically into account the local variations of background intensity.

In order to address the peculiarities of the desired segmentation process, an efficient spot detector should consider the following specifications.

- It should be insensitive to large-scale intensity variations. These are characterized by low spatial frequencies and are usually associated with the presence of mineral veins or other features of the stone.
- As the size of the spots is approximately known but may vary, the detector should be adaptable to an expected size but should not be too specific. The prior assumption about the shape of the spots is that they are round resulting to an angular isotropic operator. Thus segmented regions of linear or dot- shaped structure are not considered as deterioration patterns and, they should be eliminated by the employment of appropriate morphological operators.
- Spots of high contrast should be detected even in areas of high noise level, whereas in areas of low noise level spots of low contrast must also be detected.

To provide robust segmentation results based on the above specifications, we implemented and tested several algorithmic schemes that can be classified into different categories depending on the way they handle background inhomogeneities. The first step towards the implementation of an efficient spot detector is to decouple the detection of useful information from the background activity. This is achieved by the first algorithmic approach, which employs a broadband high-pass filter to enhance the decay areas and remove the general structure of the background. The segmentation process in this first approach is conducted through a simple thresholding technique that sets a global threshold from the statistical analysis of the entire image. The disability of such methods to eliminate the induction of false positive and false negative spots leads to the employment of the next category that uses adaptive thresholding schemes. Thus, we tested algorithmic approaches that perform thresholding based on characteristics of the local background structure using also some knowledge of the extent and spatial arrangement of decay patterns. All the above methods, however, use information from the histogram of the sub-regions in order to select an appropriate threshold. A fundamental limitation of such approaches is that they completely ignore information regarding the spatial relations of intensity values. In order to overcome this limitation, we also tested a local region growing segmentation approach. The basic goal here is to select local thresholds dynamically, based on an iterative evaluation of the labeling quality, achieved by each threshold value. At each iteration, the initially selected area is grown according to a thresholding similarity predicate aiming at producing compact areas, while avoiding the merging of different regions. In an effort to further reduce the segmentation errors, introduced due to the local background variations, we also implemented a more elaborate growing scheme that uses prior knowledge of the expected size of spots and the inter-spot distance. This procedure is quite reliable in detecting spot locations even in low contrast ratio between the spot and its background. However, the detected shape is distorted and the boundary of the individual spots is smoothed. In order to address the effective shape detection of decay spots, we tested a category of local morphological operators. This approach preserves the original spot shape, at the price of more false positive spots and merged spots that should be separated. In order to exploit the strength of both concepts (accurate topology detection and shape preservation), a morphological fusion algorithm was implemented, which expands the areas detected by the band-pass filtering approach up to the size derived by the morphological operators. These algorithms are briefly presented in the next section.

3. Segmentation procedures

3.1. Detection based on frequency selection and thresholding

In this section, we discuss segmentation approaches that involve frequency-selective filtering followed by thresholding. Most algorithmic schemes employ high-pass filtering to enhance the discernibility of discontinuities of the stone structure. The high-pass filtering process removes the low frequency content of the image that mostly reflects background activity. Following the filtering processes, they extract the histogram of the detail[1] image as to determine appropriate thresholds. Other algorithmic schemes induce

[1] Detail image is the image obtained through the frequency selection process.

band-pass filtering accompanied by a dual thresholding scheme. Through this process, we aim at maintaining patterns with specific frequency content while suppressing random variations associated to noise artifacts.

3.1.1. High-pass filtering algorithm

Considering the small size of the deterioration particles, a low pass filter with a wide kernel would be able to remove them, while conserving the general background of the image. Conversely, a high pass filter may be used to detect such decay areas. A high pass filtered image may be derived as the difference between the original and a low-pass filtered version of the image as:

$$f'(x,y) = f(x,y) - G_\sigma[f(x,y)]$$

where $f(x,y)$ is the original image and $G_\sigma[f(x,y)]$ and $f'(x,y)$ represent the low-pass and the high-pass versions of the image, respectively. As low pass filter, a Gaussian filter with a wide kernel can be employed. The parameters of the Gaussian filter are chosen to be suitable for the detection of the objects of interest. The size σ is chosen to be larger than the expected size of the majority of decay patterns. Here $\sigma = 2.75$ is chosen, implying that spatial variations at a scale larger than this are attenuated. The region of support of the Gaussian filter is 21×21 *pixels*. This size is selected to preserve the extent of the expected deterioration patterns. Subsequently, the high-pass filtered image is being thresholded in order to determine decay areas. The procedure of evaluating the thresholds is based on the histogram. Thus, the threshold values are determined according to the statistical Otsu approach [14].

3.1.2. Weighted difference of Gaussians (DoG) detector

The Difference of Gaussians Detector (DoG) employs a frequency selection process that performs band-pass filtering of the original image as to enhance discontinuities related to the presence of decay [15]. The entire approach consists of several steps. At first, the original image $f(x,y)$ is low-pass filtered using a Gaussian kernel with standard deviation σ equal to 4 pixels and its high-pass filtered version is obtained as:

$$f_1(x,y) = f(x,y) - G_4[f(x,y)] \tag{1}$$

The weighted difference of Gaussian filtering is based on the subtraction of one smoothed version of the resulting image $f_1(x,y)$ from another version having a different degree of smoothing. Two Gaussian kernels with different standard deviations are used to smooth the high-pass filtered version of the image. The standard deviations of the Gaussian kernels are chosen to reflect the dimension of the decay areas and their inter-particle distance. The detector operates by assigning a weight to the kernel of larger width. For the segmentation of decay areas in our application we recruit the following form of DoG.

$$f_2(x,y) = 0.8G_6[f_1(x,y)] - G_{0.25}[f_1(x,y)] \tag{2}$$

Subsequently, the standard deviation of the band-pass filtered image $f_2(x,y)$ is calculated and an initial threshold equal to k_1 times this standard deviation is applied. Then, the standard deviation is recalculated using only the pixels beyond the initial threshold. The final threshold is set as k_2 times the recalculated standard deviation. According to previous studies [16], k_1 and k_2 should belong in the range [1,3] if the standard deviation of the histogram of $f_2(x,y)$ is larger than 1. In our application these constants are selected experimentally and for the detection of black spots k_1 and k_2 are set to 2 and 3, respectively. As the Gaussian detector does not preserve the shape of the spots, this scheme provides reliable information about the location of decayed areas but not their shape. A methodology of segmenting decay patterns with their shape preserved is also implemented in this work. More specifically, a morphological detector based on 'bothat' filtering followed by twin thresholding is employed along with the DoG detector [15].

The previous global thresholding methods do not take under consideration specific features of the local background, thus inducing many false positive and false negative areas. The implementation of neighbor-based segmentation procedures that employ thresholds based on intensities of the neighboring pixels is more efficient in suppressing instances of over-segmentation. The latter mainly occur due to the dynamically varying stone structure. In the subsequent sections, we introduce three approaches of neighborhood-based segmentation that rely on stochastic hypotheses of the local intensity distributions.

3.2. Segmentation approaches based on local thresholds

Neighborhood-based threshold selection aims at exploiting local characteristics to reduce the false positive and false negative segments. These methods operate on larger segmented areas derived by global thresholding (results of algorithm in Section 3.1.1). After labeling these initial "candidate" areas, the subsequent segmentation is performed via employment of locally determined thresholds. The specific definition of label assignment that we adopt is the same as used by Hoover et al. [17] based on eight-connectivity for local label compactness.

For each label, the co-ordinates of the center of gravity are calculated and stored in a structure. Subsequently, a window of size 61 × 61 *pixels* centered at each specific set of coordinates is applied. The extent of the window is selected as to provide discernibility between the locations of the decay areas and the stone structure. In each window sub-area the histogram is extracted and thresholds are determined based on stochastic hypotheses for the distribution of local intensities. We test three methods of threshold selection that reflect different hypotheses regarding the local intensity distributions:

- Initially, we assume a normal distribution of local intensities. Thus, the mean and the standard deviation are considered as representative measures of the intensities' distribution. In this case, the threshold is determined via the mean and the standard deviation values. All pixels that satisfy the equation $p(i,j) \leqslant (Mean - 1.5 * s\tan dard\_deviation)$ are considered to comprise black spots.
- Another hypothesis on the intensities assumes non-parametric distribution. In this case, the threshold depends upon the median and the quartiles levels. The procedure followed to perform the Box Plot Thresholding involves extraction of the median, lower quartile and upper quartile of gray levels. The threshold applied for the detection of black spots is $Th_1 = Upper\_Quartile - 1.5 * Inter\_Quartile$. All pixels that satisfy the condition $p(i,j) \leqslant Th_1$ correspond to decay areas.
- Finally, in the Robust Fit Thresholding approach we assume that the local background intensities obey the normal distribution and a curve fitting approach is recruited to extract outliers (non-background segments), which depart from that normal distribution. Thus, we first extract the distribution of gray levels in each sub-region defined by the square window. Subsequently, a normal distribution is fit through robust *t*-fitting as to avoid the effects of outliers. The robust fit function uses iteratively reweighed least

squares algorithm and the weights at each iteration are calculated by applying the bisquare function to the residuals from the previous iteration. This algorithm assigns lower weights to points that do not fit well the histogram. Subsequently, the weights derived from the above procedure are stored in a vector. While traversing the vector from the head to the end element, the position of the first nonzero element corresponds to the threshold denoted as Th_2. All pixels with gray values lower than Th_2 are detected as pixels belonging to black particles.

3.3. Sub-region decomposition algorithm

The algorithmic scheme identified as "Sub-Region Decomposition Algorithm" also involves frequency-selective filtering for the acquisition of the detail image [18]. The thresholding approach, though, is applied only to susceptible sub-regions of the image, which are determined according to statistical properties of the intensity distribution. The applied thresholds also stem from the local intensities.

Since black particles are small isolated regions, they produce outliers in the intensity histogram of the detail image. Our segmentation problem can thus be reduced to that of detecting outliers. The detail image is first divided into square non-overlapping regions of extent 61 × 61 pixels; the dimensions are selected properly to provide sufficient discrimination between the background and the decay areas. In each of the decomposed subregions, the histogram is extracted and the Skewness and Kurtosis are computed as measures of the asymmetry and impulsiveness of the distribution. For a random variable X the Skewness is given by:

$$\gamma_3 = \frac{E\left[(X - E[X])^3\right]}{\left(E[(X - E[X])^2]\right)^{\frac{3}{2}}} \tag{3}$$

and an estimate of the Skewness is obtained as:

$$\hat{\gamma}_3 = \frac{\sum_{i=1}^{N}(X_i - m)^3}{(N-1)\sigma^3} \tag{4}$$

where m and σ indicate the estimates of mean and standard deviation over N observations of $X_i\,(i = 1,\ldots,N)$. Skewness is a measure of the asymmetry of the data around the sample mean.

Similarly, for a random variable X the Kurtosis is defined in terms of the tails of the distribution as:

$$\gamma_4 = \frac{E[(X - E[X])^4]}{\left(E[(X - E[X])^2]\right)^2} \tag{5}$$

and the estimate of Kurtosis is given by:

$$\hat{\gamma}_4 = \frac{\sum_{i=1}^{N}(Xi - m)^4}{(N-1)\sigma^4} - 3 \tag{6}$$

Kurtosis is a measure of how outlier prone a distribution is. If a region contains decay areas then due to their impulsive nature the symmetry of the detail's image histogram is altered. In this case, it is also evident that the tails of the distribution are heavier and hence the kurtosis assumes a quite high value. Therefore, the detection is performed by the following decision rule based on the Skewness and kurtosis values of the sub-regions' histograms:

$$\Gamma(x) = \begin{cases} 0 & \text{background if} \quad \gamma_3 \leqslant T_1 \quad \text{or} \quad \gamma_4 \leqslant T_2 \\ 1 & \text{decay spot if} \quad \gamma_3 > T_1 \quad \text{or} \quad \gamma_4 > T_2 \end{cases} \tag{7}$$

where T_1 and T_2 are experimentally determined thresholds. Once the sub-blocks containing the deterioration patterns are determined by the above test, the thresholding procedure estimates the locations where decay areas prevail by calculating the lower quartile (Q_1), upper quartile (Q_3) and inter-quartile range (denoted by R_f). Then, pixels with intensity levels lower than $Q_l - kR_f$, $k \in [1.5, 3]$ are assigned to degraded areas.

The algorithmic schemes discussed so far employ locally determined thresholds selected semi-automatically, where the algorithm developer still needs to determine some of the threshold parameters. A drawback of these approaches may be that they select thresholds based on the distribution of intensities at each sub-region while discarding information related to the spatial arrangement of intensity levels. The implementation of segmentation approaches that employ dynamically varying thresholds based on iterative evaluation of the labeling quality for each threshold value may provide a more efficient discrimination between decay areas and noise artifacts with fully automated threshold selection. Towards this direction, the Region Growing algorithm is considered next.

3.4. Region growing algorithm

The Region Growing Algorithm [19] starts with a high pass filtering of the image under consideration (as in Section 3.1.1). Subsequently, all pixels with intensity values under the median level are selected as seed pixels. A region is grown around a seed pixel by appending its 4-connected neighbors that satisfy the following condition.

$$f(i,j) \leqslant (1-t)\frac{F_{\min} + F_{\max}}{2} \tag{8}$$

where $f(i,j)$ is the pixel being checked, $F_{\max}$ and $F_{\min}$ are the current maximum and minimum intensities within the region being grown and t is a region growing tolerance parameter. The value of t is automatically derived for each segmented structure by repeating the growth with multiple values of t in the interval [0.01, 0.4]. The t-value that introduces the least distance between labeled features from one step to the following is chosen as the optimal tolerance value. The features studied are the centre of gravity of the segmented regions and their size. Thus, the algorithm determines the value of t that results in minimal change in the vector of two features with respect to the previous t value, by computing a normalized distance between consecutive feature vectors. The Region Growing Algorithm guarantees high detection accuracy, since it efficiently utilizes the local characteristics of the stone background by considering the gray value variations on the neighborhood of seed pixels.

3.5. Conditional thickening operation

This algorithm is based on a process of fusing the results segmented by both the DoG (Section 3.1.2) and Morphological detector (Section 3.1.2) and forms an accurate approach for estimating both the topology and extent of degraded regions [15]. The fusion is performed by the conditional thickening operator, denoted by $\otimes$, applied on a spot X relative to Y with the pair of structuring elements (M_{i1}, M_{i2}) as follows:

$$(M_{i1}, M_{i2}) \otimes XY = Y \cap \left(X \cup \left((M_{i1} \Theta X) \cap \left(M_{i2} \Theta X^C\right)\right)\right) \tag{9}$$

Thus, a spot X of the DoG detector is expanding up to the point that it reaches a neighboring spot or until it reaches the size of a co-located spot Y detected by the morphological operator. The pair of structuring elements M_{i1} and M_{i2} controls the direction of expansion. To cover spatial expansion in many directions, we use eight pairs of such elements. The first two pairs are given as:

$$(\mathbf{M}_{11}, \mathbf{M}_{12}) = \left(\begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix}\right) \text{ and } (\mathbf{M}_{21}, \mathbf{M}_{22}) = \left(\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix}\right)$$

The remaining pairs are obtained from these matrix combinations through rotations every 90°. Finally, the conditional thickening operator is obtained as a combination of individual results for every pair (M_{i1}, M_{i2}) as:

$$E = \bigcup_{i=1}^{8} (M_{i1} M_{i2}) \otimes XY \tag{10}$$

The operator in Eq. (10) forms the core of the detection approach that employs the conditional thickening operator to combine and fuse the results of two individual detectors. For the decay areas, the patterns detected by the Gaussian detector are extended in space but the result is always intersected with the spots detected by the morphological detector. The intersection in each step preserves only spots that are collocated in both X (after conditional thickening) and Y [15].

3.6. Ground truth extraction procedure

Further to detecting degraded patterns on stone surfaces, this paper also aims at assessing the potential and the limitations of each algorithmic scheme. Towards this direction, the determination of a Ground Truth Matrix of decay regions is critical, as it provides a test bed for measuring the algorithms' performance and comprehending the differences among them on the segmentation procedure. In this work, we introduce a semi-automated approach of extracting the Ground Truth Matrix. We first extract the decay areas consistently detected by all algorithmic schemes examined. This process is considered in Section 3.6. Furthermore, the entire result is supervised and evaluated by the experts as explained in Section 4.1.

Through the Ground Truth Extraction Approach, we check in pairs the areas segmented by all algorithms. The procedure starts by labeling the segments detected by each algorithmic approach. For a pair of segmented and labeled images, Fig. 1 illustrates the processes of managing the non-overlapping and partially overlapping labels. The various steps are presented in the following subsections.

This scheme is applied on the segmented images by each and every algorithm in an incremental way, as to extract the Ground Truth Matrix of decay areas. In fact, at the ith step of the process ($i > 1$) the input images involve the result of the algorithm A_i

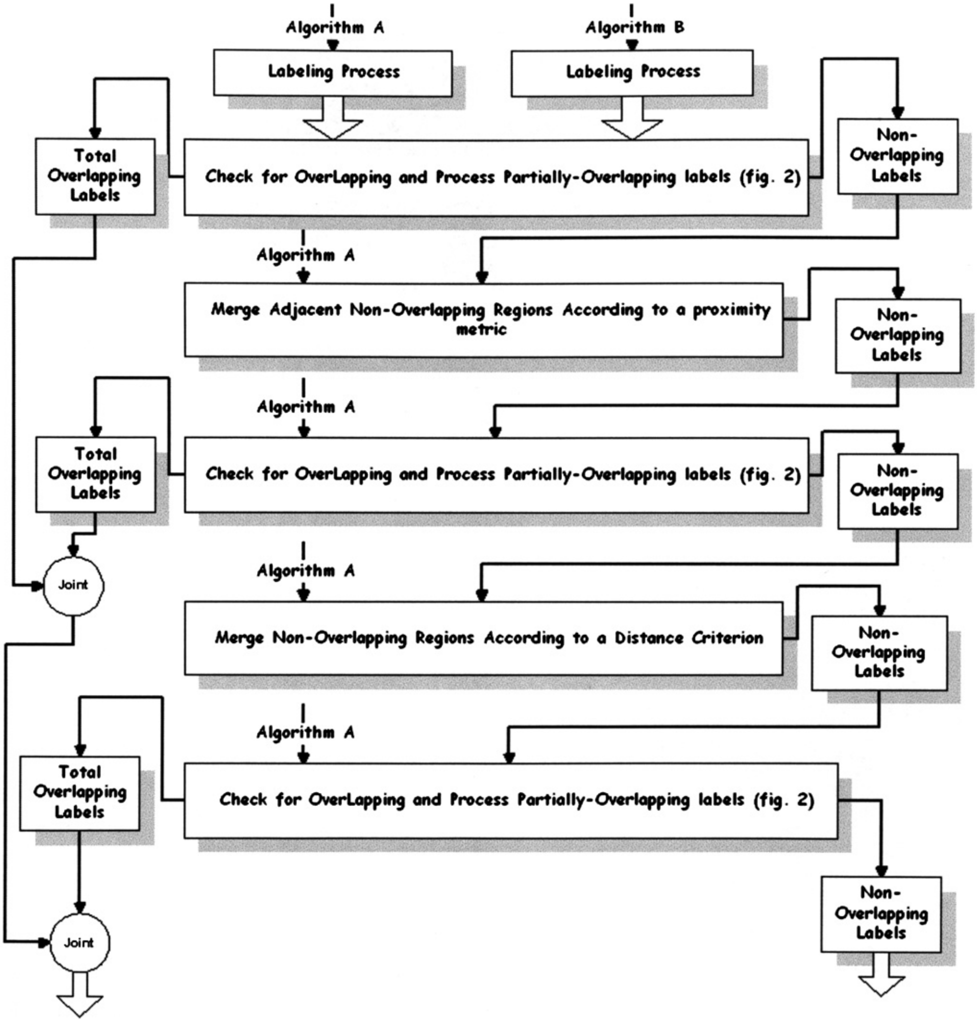
Algorithm A
Algorithm B
Labeling Process
Labeling Process
Total Overlapping Labels
Check for OverLapping and Process Partially-Overlapping labels (fig. 2)
Non-Overlapping Labels
Algorithm A
Merge Adjacent Non-Overlapping Regions According to a proximity metric
Non-Overlapping Labels
Algorithm A
Total Overlapping Labels
Check for OverLapping and Process Partially-Overlapping labels (fig. 2)
Non-Overlapping Labels
Joint
Algorithm A
Merge Non-Overlapping Regions According to a Distance Criterion
Non-Overlapping Labels
Algorithm A
Total Overlapping Labels
Check for OverLapping and Process Partially-Overlapping labels (fig. 2)
Non-Overlapping Labels
Joint

Fig. 1. Flowchart illustrating the overlap extraction procedure.

and the image representing the non-overlapping labels obtained from the $(i-1)$th step of this process.

3.6.1. Checking for overlapping labels

The Checking for Overlap aims at extracting overlapping regions detected by two different algorithms. The process initially checks whether a label detected by algorithm A_i is also segmented by algorithm A_j. The areas derived by the Checking for Overlap step can be sub-divided into three clusters. A label of A_i that does not overlap any label detected by A_j is identified as Non-Overlapping label, while labels of A_i that either in part or fully overlap labels of A_j are considered as partially and totally overlapping labels, respectively. At this point we should make clear that the aim of the Ground Truth Matrix Extraction is to mark compact areas that correspond to susceptible degraded regions. Totally overlapping labels are included in the Ground Truth. Visual inspection reveals that the partially overlapping labels often correspond to larger in extent regions that became split. In an attempt to segment the degraded regions as compact areas that represent decay patterns at their actual size, we further process the partially overlapping patterns of A_i to attain total overlap to the labels of A_j.

3.6.2. Processing the partially overlapping labels

Through this procedure we consider the partially overlapping labels of A_i that are obtained by the above process (3.6.1), in combination with the areas segmented by algorithm A_j. Initially, the partially overlapping labels of algorithm A_i are blown via a conditional thickening operator up to the point that they cover the entire corresponding label segmented by A_j. The operator of thickening label A_i using a pair of structuring elements E_1, E_2 is defined similar to Eq. (9) as:

$$(E_1, E_2) \otimes A_i A_j = A_i \cap \left(A_j \cup \left((E_1 \Theta A_j) \cap \left(E_2 \Theta A_j^c\right)\right)\right) \tag{11}$$

The segmented areas derived after processing the partially overlapping patterns are labeled and treated similar to the total overlapping labels. Fig. 2 illustrates the algorithmic procedure described above. The next step of the Ground Truth extraction approach involves processing the Non-Overlapping Labels. The Ground Truth Matrix includes all labels segmented by all or just some of the algorithms, recognizing the potential of an algorithm's failure in spot detection. Thus, even non-overlapping patterns between two algorithms may actually be part of the Ground Truth of the problem.

3.6.3. Processing the non-overlapping patterns

The majority of non-overlapping labels correspond to areas small in extent appearing in clusters, which may reflect the algorithms' inability to reveal the entire extent of decay areas, detecting only broken regions. To overcome these instances of over-segmentation,

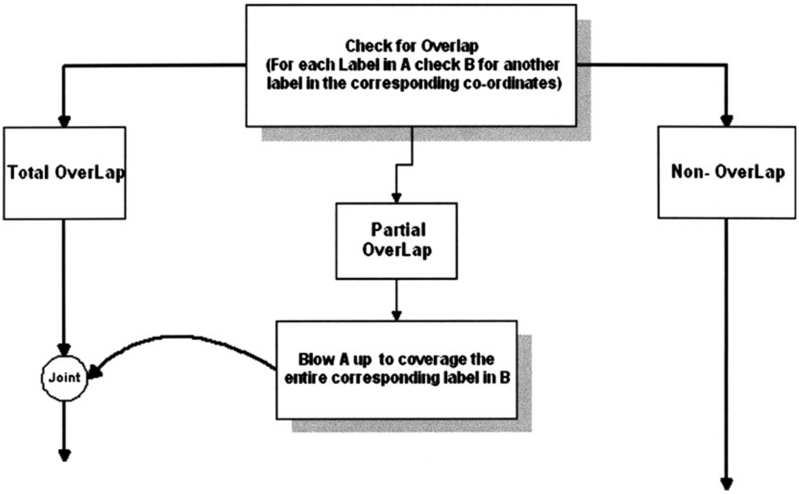
Check for Overlap
(For each Label in A check B for another label in the corresponding co-ordinates)
Total OverLap
Partial OverLap
Non- OverLap
Blow A up to coverage the entire corresponding label in B
Joint

Fig. 2. Check for overlap spots and processing the partially overlapping labels.

a process of merging adjacent non-overlapping spots is developed. We initially start by merging such areas detected in close distances.

3.6.3.1. Merging adjacent non-overlapping areas. The procedure starts by labeling all non-overlapping areas and extracting the centroid of each label. Subsequently, we measure the Euclidean distance between a label's centroid to the centroids of its neighboring labels and if the distance is lower than a predefined threshold T_D, then the adjacent labels are merged. In the current implementation, the distance threshold T_D is chosen to reflect the mean diameter value of all overlapping spots. After the process of merging the neighboring patterns, the areas obtained are labeled again and the procedure illustrated in Fig. 2 is repeated to check for new overlap. The next step involves measuring the cross-image distance between the new non-overlapping areas and any areas detected by A_j at adjacent locations.

3.6.3.2. Merging according to a distance criterion. The non-overlapping labels provided by the previous step are labeled and the centroid of each label is extracted. A window of size 31×31 is applied both at the centroid of the label in image A_j and at the corresponding coordinates in image A_i. The label in the window defined in A_i is submitted to morphological erosion (Dilates black spots) by a structuring element (disk). The morphological erosion is iterated by increasing the radius by 1 at each iteration and terminates when either the label in window A_i overlaps an existent label in window A_j or the radius value reaches an upper bound. This procedure is repeated for each of the segmented areas and the radius values at which the morphological operation terminates are used to calculate the median erosion value. Subsequently, morphological erosion is performed on all original areas of A_i with a disk-structuring element of the size of the median. Finally, the process in Fig. 2 is applied again, to derive the new overlapping labels.

3.7. Evaluation measures

3.7.1. Receiver operating characteristic curves (ROC)

The segmentation of an image through an algorithmic approach, referred to as algorithmic segmentation (AS), is compared with the corresponding ground truth (GT) specification as to account for instances of correct segmentation, under-segmentation, over-segmentation, missed regions, and noise regions. The definitions of metrics are based on the determination of overlap in terms of pixels commonly segmented in AS and GT. Based on this comparison, we compute the instances of false positive (FP), false negative (FN), true positive (TP) and true negative (TN) segments.

In an attempt to illustrate the algorithms' performances and their differences associated to the segmentation procedure, the receiver operating characteristic (ROC) curves are constructed [20–22]. The ROC curves are obtained by modifying the thresholds within meaningful ranges and subsequently calculating instances of correct and incorrect segmentation. As the thresholds are varied from lower (more strict) to higher (relaxed) values, the number of instances of correct segmentation increases but the sensitivity of the algorithms is reduced.

3.7.2. Tests of statistical significance

Further to evaluating the algorithms' performances, we are also interested in investigating the significance of variations of decay patterns that are caused by the different natural conditions of exposure. Similar methodology can be followed in assessing differences in the effectiveness of cleaning methods. The significance of such variations on the size of decay patterns is assessed through the Mann–Whitney *U* test, while intensity variations are considered through *t* tests. For the former, we select a non-parametric rank sum test, since the distribution of sizes departs significantly from the normal distribution. Regarding intensity comparisons, the *t* tests are employed after assessing that intensities on decay areas obey the normal distribution. In order to increase the size of the test set, each of the studied images is divided into six sub-blocks of equal size and the tests are performed in the sub-blocks areas.

3.7.2.1. t Tests [23,24]. Through the *t* test, the statistical parameters in concern include the mean intensity and its standard deviation for each population tested. Test sets are obtained for each sub-block of the same type and all segmented areas are subsequently used to calculate the *t*-statistic:

$$t = \frac{M_1 - M_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{12}$$

where s is an estimate of the standard deviation based on both sample populations and n_1 and n_2 are the numbers of observations within each group. Thus,

$$s = \sqrt{\frac{(n_1 - 1) \times \mathrm{std}_1 + (n_2 - 1) \times \mathrm{std}_2}{n_1 + n_2 - 2}} \tag{13}$$

where std_1 and std_2 denote the standard deviations measured in the populations 1 and 2, respectively, with

$$\text{degrees of freedom(df)} = n_1 + n_2 - 2 \tag{14}$$

3.7.2.2. Mann–Whitney U test [23,25]. The implementation of the Mann–Whitney *U* test proceeds as follows, given two population groups:

1. List the observations in order of magnitude within each group. Assign ascending ranks to the entire set of observations with repeated values, called ‘ties’, given the mean of the ranks within that run.
2. Sum the ranks of each population $R_A R_B$.
3. Calculate U_A and U_B, e.g., $U_A = \{n_A(n_A + 1)/2 + (n_A n_B)\}$-$\mathrm{R_A}$ where n_A and n_B are the number of samples in each group; U_B is similarly computed.
4. Enter the smallest of U_A and U_B to the statistical table. Values of U lower than the tabulated value of significance indicate significant differences between the populations.

4. Results

This work initially validates the potential and the limitations of each of the recruited algorithms in effectively determining the topology and extent of decay patterns. A second objective is to study the size and relative intensities (over the background) of degraded areas as representative measures of the severity of degradation. In order to assess whether significant differences occur between decay patterns segmented on various surfaces, statistical tests are first employed on intensity distributions to examine the darkness of corroded areas over the background. Furthermore, tests of statistical significance are also used to evaluate differences on decay patterns' sizes associated to the exposure conditions or the cleaning state of the stone material. As a side effect, these tests contribute to comprehending the mechanisms and the efficiency of chemical cleaning, as well as to understanding the formation of crusts.

4.1. Visual evaluation of the ground truth matrixes

The images used to extract the Ground Truth Matrix are selected by the experts as to closely represent the surfaces studied throughout this work. More specifically, we evaluate the Ground Truth on three surfaces: (a) an untreated sheltered fluting monitored by FOM (Fig. 3), (b) an untreated unsheltered fluting also monitored by FOM (Fig. 4) and (c) a stone surface depicting the successive co-existence of treated and untreated strips monitored by a Digital Camera (Fig. 5). In any case, the number of decay patterns detected

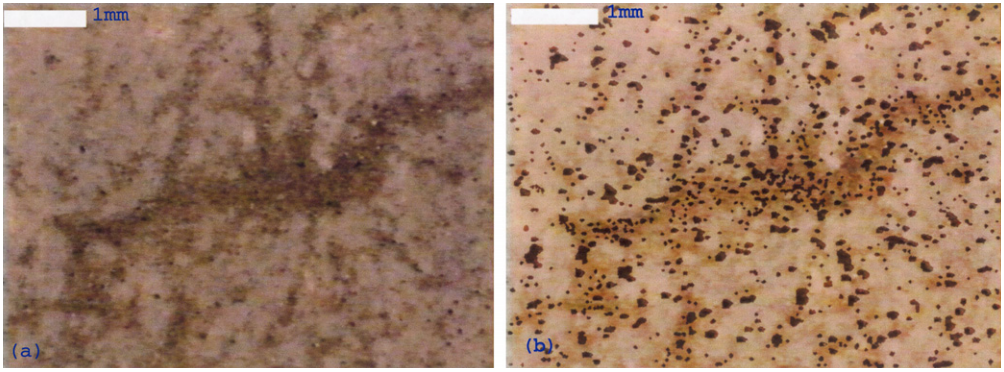

Fig. 3. (a) Stone specimen located on a column's fluting at sheltered surface (as monitored by the FOM (magnification ×50)), (b) the derived Ground truth Matrix overlaid on the original image.

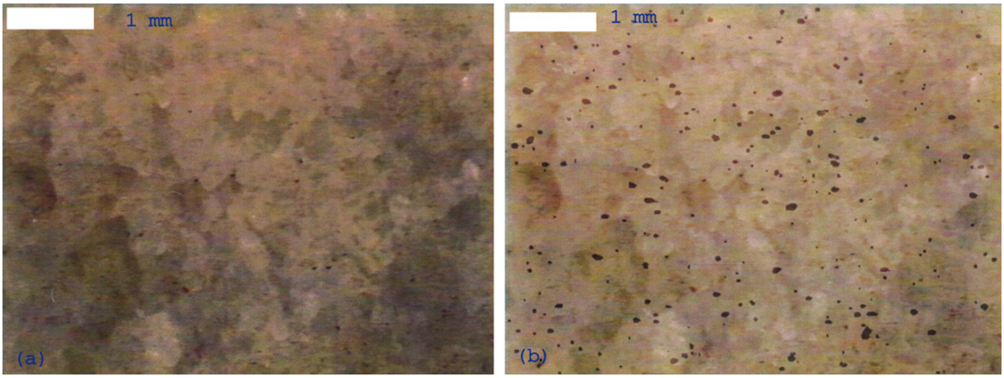

Fig. 4. (a) Stone specimen located on column's fluting at unsheltered surface (as it was monitored by the FOM system (magnification ×50)), (b) the derived Ground truth Matrix overlaid on the original image.

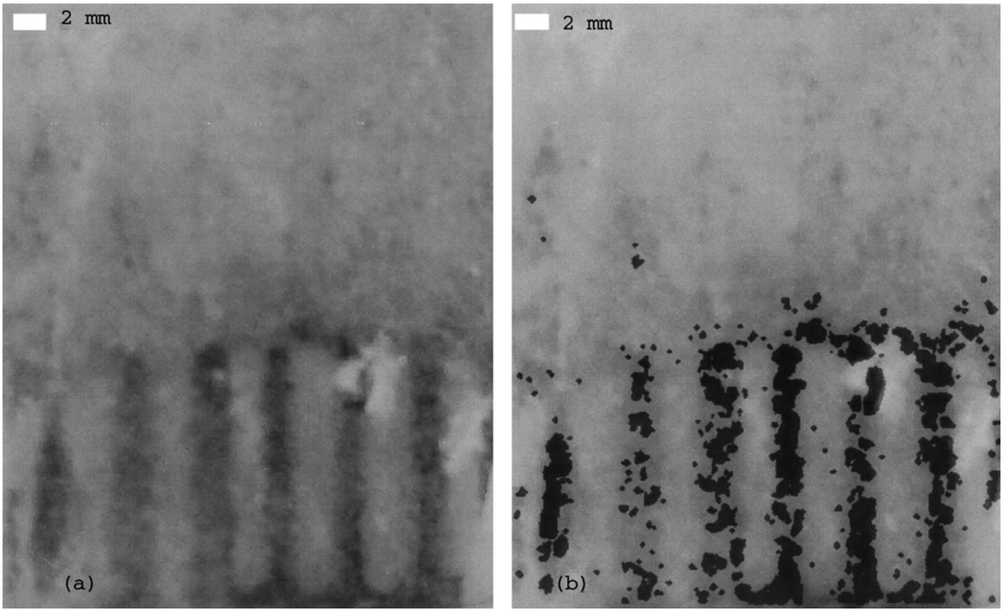

Fig. 5. (a) Stone surface monitored by the digital camera, (b) the derived Ground truth Matrix overlaid on the original image.

in each image is quite large as to form a quite valid statistical set for algorithmic comparisons As mentioned before, the extraction and validation of GT matrix is performed in two steps: (i) extraction of GT matrix based on the algorithmic results of several algorithms (Section 3.3), and (ii) validation of the GT matrix through inspection by experts. Fig. 3(a) depicts a surface of rapidly varying stone structure (untreated sheltered fluting).

According to the experts' judgment, the GT approaches quite effectively the topology of degraded areas. Moreover, the segmented corrosion patterns are large in extent marking extensive susceptible areas. Fig. 4 illustrates the ground truth matrix extracted on an image depicting an unsheltered untreated fluting.

Observing in parallel the GTs illustrated in Figs. 3 and 4 we may assess that the segmented areas are larger in extent in the case of the sheltered untreated fluting (Fig. 3). This conclusion is in accordance with the experts' judgment concerning the corrosion state encountered on sheltered and unsheltered surfaces. Finally, in order to provide a visual inspection of the Ground Truth matrix for the case of digital camera images we illustrate in Fig. 5 the original image as well as the Ground Truth Matrix.

Fig. 5(a) depicts a stone surface partially cleaned by a Nd:YAG laser cleaning. In the stone material, we observe the co-occurrence of successive cleaned and un-cleaned strips. According to the experts' estimation, the ground truth matrix has effectively determined the presence of degradation particles.

At this point, we should make clear that the objective of our detection processes is not to segment areas of intensity alteration induced by corrosion damage, but rather to determine the individual decay patterns appearing within any background structure (corroded or cleaned), which lead to the formation of black crusts beyond the color alteration effects. This applies especially for Fig. 5 where segmentation does not aim to distinguish cleaned from corroded areas but to detect decay patterns on each of these areas. The presence of small in extent regions is limited in the GT image in Fig. 5(b). This is explained by the fact that the digital camera provides low-resolution levels and thus the segmentation procedures are mainly based on large scale intensity alterations and do not require high sensitivity or adaptation.

4.2. Evaluating the algorithms' performance through ROC curves

In this paper, we consider the ROC curves as robust measures for evaluating the algorithms' performance. Throughout this subsection, we thoroughly discuss the performance curves derived for each of the GTs separately.

4.2.1. Algorithms' performance on unsheltered untreated fluting

Initially we study the algorithms' performance in the case of the unsheltered untreated fluting. From Fig. 6 it can be concluded that the Conditional Thickening Algorithm demonstrates better performance (top curve) in detecting decay patterns at their real extent, while the Region Growing follows in performance. At this point it should also become clear that the above approach of determining the specificity and sensitivity of algorithms

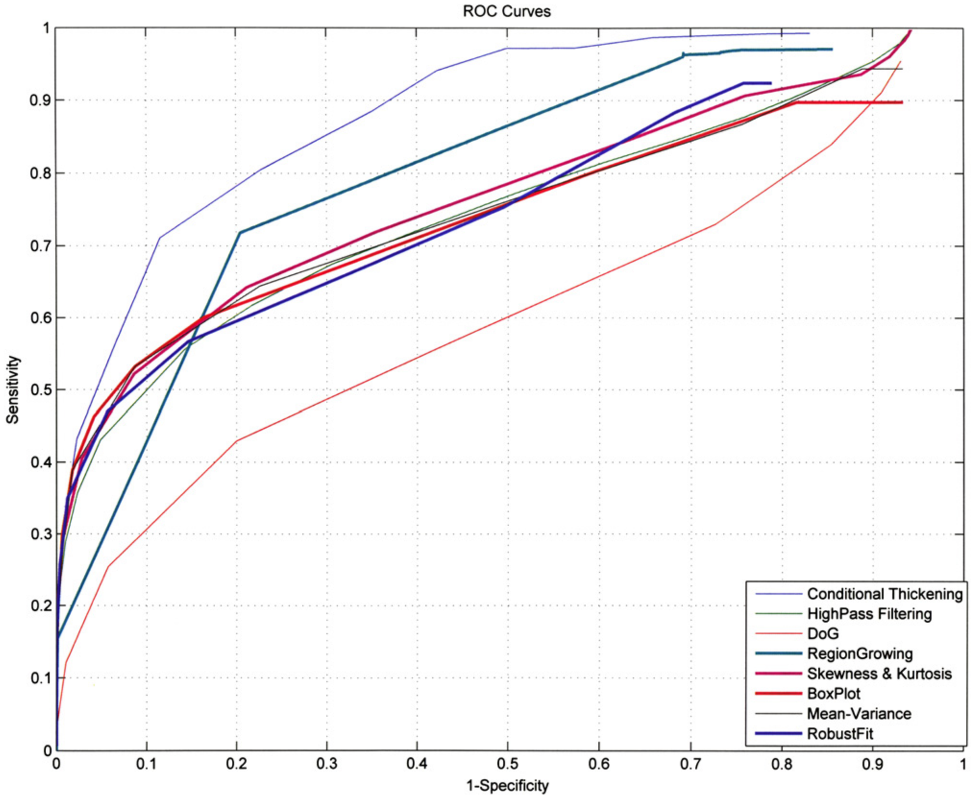

Fig. 6. ROC curves depicting the performance of the implemented algorithms in the case of the unsheltered untreated fluting (illustrated in Fig. 4).

is more focused towards the accurate detection of patterns' size, since the comparison of results with the GT targets exactly this aspect. Thus, it is expected that approaches with tendency of splitting the decayed areas, such as the methods of Adaptive Thresholding (Section 3.2), Sub-Region Decomposition and DoG, will demonstrate worse performance. By observing Fig. 6 we can also deduct that the adaptive thresholding algorithmic schemes (Section 3.1.3) tend to perform better than the High pass Filtering Algorithm for low values of sensitivity. This means that the former introduce less FP and FN areas when more strict thresholds are applied. Thus, it further reflects the potential of adaptive thresholding techniques in accurately segmenting decay spots in non-homogeneous background. A further assessment that can be drawn by the above figure is that the DoG detector appears to be inefficient. This is explained by its tendency to split the detected areas resulting in the segmentation of many spots reduced in size. Finally, regarding the Region Growing Algorithm, a remarkable point is that for *specificity values* >0.85 (1 − *specificity* <0.15) it seems less efficient than the other techniques, while for specificity values <0.85 its performance becomes better. Visual inspection of the segmentation results reveal that FPs generally correspond to small in extent and isolated areas. Their spatial arrangement is assessed by the authors, but not presented here, through measuring the mean of the minimum Euclidean inter-particle distance.

4.2.2. Algorithms' performance on sheltered untreated fluting

In this section, we illustrate the algorithms' performance in the case of a surface demonstrating a rapidly varying background structure. Fig. 7 depicts the algorithms' performance through the ROC curves.

Fig. 7 reveals that the Conditional Thickening and the Region Growing algorithms perform better than the others for specificity levels <0.35. The adaptive thresholding schemes appear to have similar responses as the high-pass filtering and the sub-region decomposition algorithms. The DoG detector, though, demonstrates a consistent worse performance. In an effort to compare the algorithms' performances as they are depicted in Figs. 6 and 7, we can see that all the algorithms, except for the Conditional Thickening and the Region Growing Algorithms, seem to be more efficient when applied to images depicting texture variations (Fig. 7). This effect can be explained by considering that decay spots on homogeneous surfaces usually correspond to areas of low contrast to the background while the opposite occurs on in-homogeneous surfaces. Thus, for the latter case the topology of decay patterns may be approached even by effective strict thresholding schemes. Regarding the Conditional Thickening and the Region Growing Algorithms we can state that they respond in almost the same way when applied to surfaces depicting homogeneous and in-homogeneous structures. Both algorithms demonstrate a slightly better performance in the case of inhomogeneous background, for specificity values around the mean of the range. This ability to provide similar accuracy when handling surfaces of different texture characteristics reflects their potential to perform efficient detection irrespective of noise levels and variations over the background.

Another important point is the response of the Adaptive Thresholding schemes (Section 3.1.3). It has been shown that for small values of specificity, these algorithms perform better on the sheltered untreated flutings (Fig. 7) compared to the unsheltered ones (Fig. 6). This is expected because texture in-homogeneities induce outliers on the histogram of the studied surfaces. Thus, susceptible areas can be segmented even by less strict thresholds. In contrast, when operating on images of smoother background, the

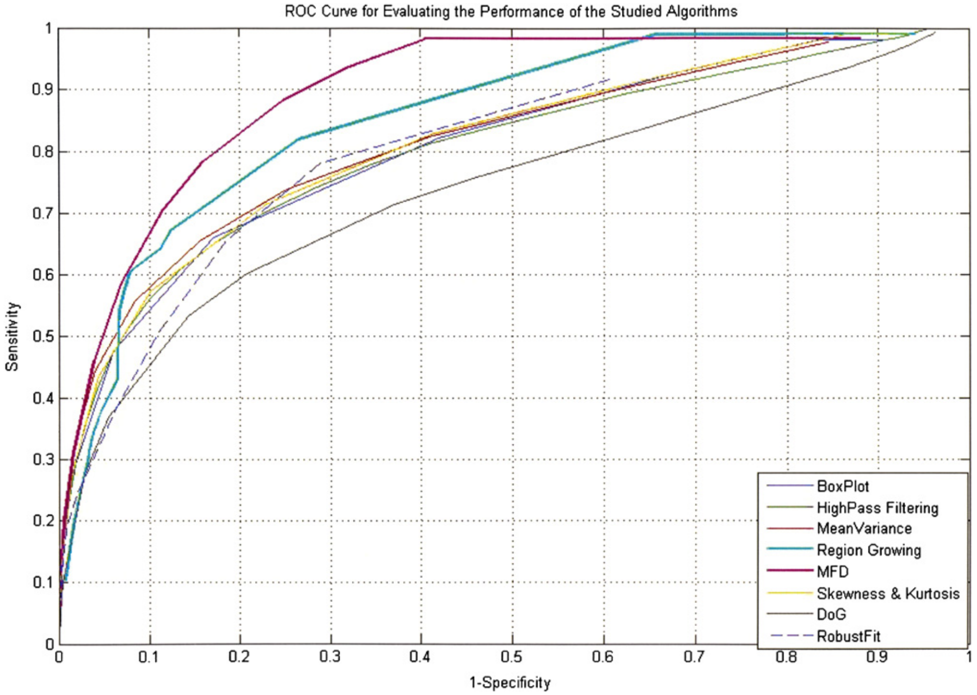

Fig. 7. ROC curves depicting the performance of the implemented algorithms in the case of the sheltered untreated fluting (illustrated in Fig. 3).

adjustment of low thresholds simply causes the segmentation of large compact areas that do not always correspond to susceptible regions.

4.2.3. Algorithms' performance on digital camera images

Further to validating the algorithms' efficiency in segmenting decay areas on FOM images, we also investigate their potential in determining corrosion effects on surfaces screened by other imaging modalities (digital camera system). Such responses are also evaluated through the ROC curves.

In Fig. 8 we can observe that the HighPass Filtering Algorithm can detect decay effects quite effectively. This is expected considering the low-resolution provided by the digital camera. In this case of low detail, a global processing algorithm can provide quite accurate results. The Conditional Thickening and the Region Growing algorithms demonstrate poorer performance than the High-Pass Filtering. Furthermore we can observe that all the algorithms' performances tend to converge for specificity <0.5.

4.3. Evaluation of the effects of structural and cleaning conditions

The design structure of artwork has direct implications to the exposure of its various surfaces (flutings, readings, etc.) to environmental conditions and also affects the efficiency of cleaning. Furthermore, different cleaning methods differ in their efficiency of removing degradation from such surfaces, but the effectiveness of cleaning is very difficult to be

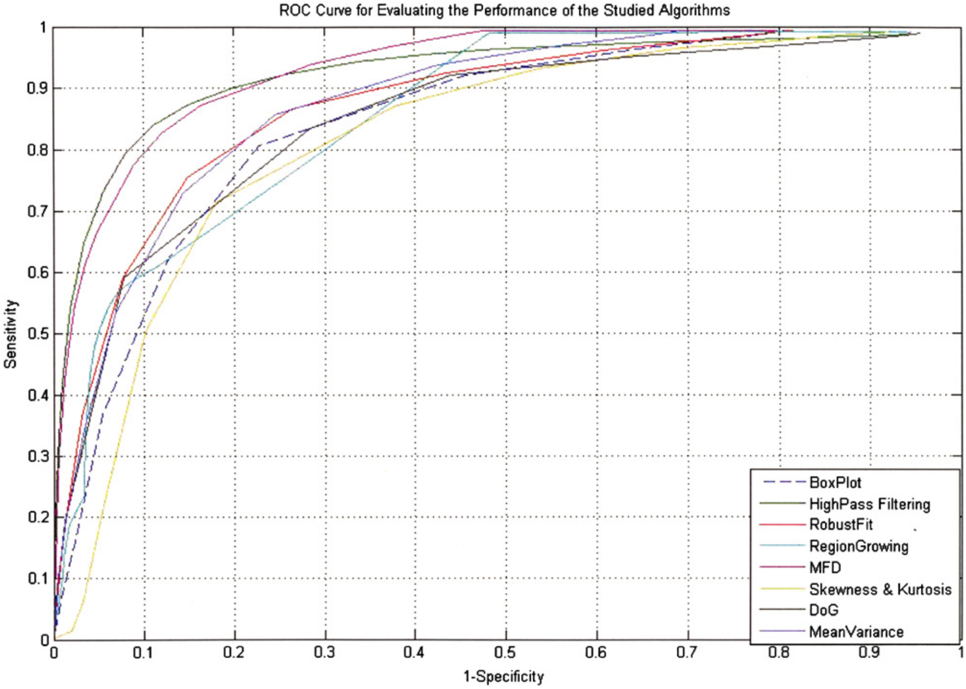

Fig. 8. ROC curves depicting the performance of the implemented algorithms for the stone material monitored via the digital camera (illustrated in Fig. 5).

quantified and compared by just human inspection. In this section, we investigate the effects of exposure and/or cleaning conditions on the size and the relative intensities (over the background) of the detected corrosion defects. After the segmentation process, two features, i.e. the extent and relative intensity of each decay region are measured. The statistical tests employed, aim at assessing whether or not decay patterns encountered on surfaces of various exposure and/or cleaning states are characterized by feature values belonging to different populations. Through this analysis, we recruit the three segmentation algorithms that perform best (according to the ROC curves), namely the Conditional Thickening, the Region Growing and the Sub-Region Decomposition algorithm.

4.3.1. *t Tests for intensity distributions*

We introduce *t* tests to estimate whether the cleaning state and/or the exposure conditions of the stone material are indeed reflected on the relative intensities (over the background) of the corroded areas. According to the experts, the darkness at the locations of decay particles is closely related to the crusts' thickness on these areas, as crusts of greater thickness absorb larger amounts of illumination. An accurate metric of darkness is the relative intensity of pixels within degraded areas over the background. Prior to the application of the *t* tests the set of images is submitted to intensity normalization to eliminate the effects of different luminance conditions. Table 1 presents and compares the results after the application of *t* tests on the studied surfaces; the results of the chemical analysis of the studied surfaces are also reported.

Table 1
Comparative study on the significance of intensity alterations

		Conditional thickening	Sub-region decomposition	Region growing
1.	Sheltered flutings (Ds) (vs.) Sheltered flutings (Diagn.)	Df = 34 *t* = 25.764 Critical *t* (1-tail) = 1.691	Df = 34 *t* = 22.187 Critical *t* (1-tail) = 1.691	Df = 34 *t* = 24.478 Cr.t(1-tail) = 1.691
2.	Sheltered flutings (WMB) (vs.) Sheltered flutings (Diagn.)	Df = 33 *t* = 62.410 Critical *t* (1-tail) = 1.692	Df = 32 *T* = 63.829 Critical *t* (1-tail)= 1.694	Df = 32 *t* = 59.279 Critical *t* (1-tail)= 1.694
3.	Sheltered flutings (BP) (vs.) Sheltered flutings (Diagn.)	Df = 27 *t* = 33.899 Critical *t* (1-tail) = 1.703	Df = 28 *t* = 40.619 Critical *t* (1-tail) = 1.701	Df = 28 *t* = 36.787 Critical *t* (1-tail) = 1.701
4.	Sheltered reedings (Ds) (vs.) Sheltered reedings (Diagn.)	Df = 9 *t* = 12.591 Critical *t* (1-tail) = 1.833	Df = 10 *t* = 14.021 Critical *t* (1-tail) = 1.812	Df = 10 *t* = 13.712 Critical *t* (1-tail) = 1.812
5.	Sheltered reedings (BP) (vs.) Sheltered reedings (Diagn.)	Df = 8 *t* = 12.716 Critical *t* (1-tail)= 1.860	Df = 10 *t* = 18.321 Critical *t* (1-tail) = 1.812	Df = 10 *t* = 14.436 Critical *t* (1-tail) = 1.812
6.	Sheltered reedings (Diagn.) (vs.) Sheltered flutings (Diagn.)	Df = 28 *t* = 13.443 Critical *t* (1-tail) = 1.701	Df = 22 *t* = 13.388 Critical *t* (1-tail) = 1.717	Df = 28 *t* = 10.148 Critical *t* (1-tail) = 1.701
7.	Unsheltered flutings (Diagn.) (vs.) Sheltered flutings (Diagn.)	Df = 34 *t* = 47.960 Critical *t* (1-tail) = 1.691	Df = 34 *t* = 31.016 Critical *t* (1-tail) = 1.691	Df = 34 *t* = 4 1.429 Critical *t* (1-tail) = 1.691
8.	Unsheltered flutings (DS) (vs.) Unsheltered flutings (Diagn.)	Df = 22 *t* = 7.749 Critical *t* (1-tail) = 1.717	Df = 22 *t* = 12.089 Critical *t* (1-tail) = 1.717	Df = 22 *t* = 10.765 Critical *t* (1-tail) = 1.717
9.	Unsheltered flutings (Diagn.) (vs.) Sheltered reedings (Diagn.)	Df = 16 *t* = 16.347 Critical *t* (1-tail) = 1.746	Df = 16 *t* = 18.940 Critical *t* (1-tail) = 1.746	Df = 16 *t* = 16.487 Critical *t* (1-tail) = 1.746
10.	Unsheltered reedings (Diagn.) (vs.) Unsheltered flutings (Diagn.)	Df = 22 *t* = 4.575 Critical *t* (1-tail) = 1.717	Df = 22 *t* = 6.799 Critical *t* (1-tail) = 1.717	Df = 22 *t* = 4.090 Critical *t* (1-tail) = 1.717
11.	Unsheltered reedings (Ds) (vs.) Unsheltered reedings (Diagn.)	Df = 9 *t* = 7.576 Critical *t* (1-tail) = 1.833	Df = 10 *t* = 8.401 Critical *t* (1-tail) = 1.812	Df = 10 *t* = 7.004 Critical *t* (1-tail) = 1.812
12.	Unsheltered reedings (WMB) (vs.) Unsheltered reedings (Diagn.)	Df = 9 *t* = 8.42 Critical *t* (1-tail) = 1.833	Df = 10 1 = 11.63 Critical *t* (1-tail) = 1.812	Df = 10 *t* = 7.770 Critical *t* (1-tail) = 1.812

Table 1 (*continued*)

		Conditional thickening	Sub-region decomposition	Region growing
13.	Unsheltered reedings (Diagn.) (vs.) Sheltered reedings (Diagn.)	Df = 16 $t = 25.223$ Critical t (1-tail) = 1.746	Df = 16 $t = 30.958$ Critical t (1-tail) = 1.746	Df = 16 $t = 15.972$ Critical t (1-tail) = 1.746

According to the statistics setup of the test, t-values that are greater than the critical 't' value reflect a corresponding difference in the mean values of the examined populations. Through the tested hypotheses, we aim at assessing the occurrence of differences between the studied samples. As null hypotheses we always state that the first of the populations has a distribution of intensity values laid on lower levels, while the alternative hypotheses state the opposite. To perform such a test, where the rejection region corresponds to the largest values, we employ one-tailed statistical test. When the t-value is much larger than the critical t, then the null hypothesis is rejected in favor of the alternative hypothesis of different distributions.

The conclusions drawn through the data presented in Table 1 reveal that the application of cleaning interventions results in increasing the intensity levels of the remaining corroded areas. This effect holds true for all cleaning methods. In particular, the corroded areas detected on surfaces cleaned by DS appear to be darker than decay spots segmented on surfaces treated by other cleaning processes (Table 1). This finding is in accordance with the results of the chemical analyses indicating that the DS cleaned areas still contain aluminosilicates and gypsum relevant to the presence of black particles (see Table 1). This has also theoretical relevance, since most sheltered surfaces are associated with higher amounts of decay products, such as gypsum, aluminosilicates, nitrates and organic compounds. A similar effect is observed when comparing surface segments at different structural position of the stonework. More specifically, decay patterns segmented on sheletered flutings are darker than the corresponding patterns detected on sheltered reedings. An effort to investigate whether a similar observation is also valid for the unsheleterd areas revealed that the observed difference on the relative intensity values, is only marginally significant. This conclusion also agrees to the chemical analysis.

Through Table 1 we can draw important conclusions regarding the algorithmic responses. Thus, we can observe that the sub-Region Decomposition algorithm provides results demonstrating discrepancies from the results derived by the other two algorithms. This is mainly caused by the fact that it tends to split areas segmented as compact by the other algorithms. Thus, such a behavior affects the distribution of intensities.

4.3.2. Mann–Whitney U test for size distributions

Through the Mann–Whitney U test we aim at investigating whether an association between the corrosion state and the size of decay patterns can be established. The employment of the specific statistical test was decided because, according to our observations, the distribution of segments sizes departs significantly from the normal distribution. Table 2 summarizes the results of the Mann–Whitney U test when applied on the studied surfaces.

The hypotheses tested in this approach are quite similar to those tested in the t tests. Thus, at null hypotheses we assume that decay areas belonging to the first population

Table 2
Comparative study on the significance of decay patterns size alterations

		Conditional thickening	Region growing	Sub-region decomposition
1.	Sheltered flutings (Diagn.) (vs.) Sheltered flutings (Ds)	$N_1 = 24$, $N_2 = 12$ $U = 252$, $U_{\text{crit}} = 74$ $p = 5.92 \times 10^{-5}$	$N_1 = 24$, $N_2 = 12$ $U = 288$, $U_{\text{crit}} = 74$ $p = 7.98 \times 10^{-10}$	$N_1 = 24$, $N_2 = 12$ $U = 288$, $U_{\text{crit}} = 74$ $p = 7.98 \times 10^{-10}$
2.	Sheltered flutings (Diagn.) (vs.) Sheltered flutings (WMB)	$N_1 = 24$, $N_2 = 6$ $U = 144$, $U_{\text{crit}} = 27$ $p = 1.68 \times 10^{-6}$	$N_1 = 24$, $N_2 = 10$ $U = 288$, $U_{\text{crit}} = 74$ $p = 7.98 \times 10^{-10}$	$N_1 = 24$, $N_2 = 18$ $U = 432$, $U_{\text{crit}} = 124$ $p = 10^{-6}$
3.	Sheltered flutings (Diagn.) (vs.) Sheltered flutings (BP)	$N_1 = 24$, $N_2 = 6$ $U = 144$, $U_{\text{crit}} = 27$ $p = 1.68 \times 10^{-6}$	$N_1 = 24$, $N_2 = 6$ $U = 144$, $U_{\text{crit}} = 27$ $p = 1.68 \times 10^{-6}$	$N_1 = 24$, $N_2 = 6$ $U = 144$, $U_{\text{crit}} = 27$ $p = 1.68 \times 10^{-6}$
4.	Sheltered reedings (Diagn.) (vs.) Sheltered reedings (DS)	$N_1 = 6$, $N_2 = 6$ $U = 36$, $U_{\text{crit}} = 3$ $p = 10.8 \times 10^{-4}$	$N_1 = 6$, $N_2 = 6$ $U = 36$, $U_{\text{crit}} = 3$ $p = 10.8 \times 10^{-4}$	$N_1 = 6$, $N_2 = 6$ $U = 36$, $U_{\text{crit}} = 3$ $p = 10.8 \times 10^{-4}$
5.	Sheltered reedings (Diagn.) (vs.) Sheltered reedings (BP)	$N_1 = 6$, $N_2 = 6$ $U = 36$, $U_{\text{crit}} = 3$ $p = 10.8 \times 10^{-4}$	$N_1 = 6$, $N_2 = 6$ $U = 36$, $U_{\text{crit}} = 3$ $p = 10.8 \times 10^{-4}$	$N_1 = 6$, $N_2 = 6$ $U = 36$, $U_{\text{crit}} = 3$ $p = 10.8 \times 10^{-4}$
6.	Sheltered flutings (Diagn.) (vs.) Sheltered reedings (Diagn.)	$N_1 = 24$, $N_2 = 6$ $U = 0$, $U_{\text{crit}} = 27$ $p = 1.68 \times 10^{-6}$	$N_1 = 24$, $N_2 = 6$ $U = 18$, $U_{\text{crit}} = 27$ $p = 16.7 \times 10^{-4}$	$N_1 = 24$, $N_2 = 6$ $U = 18$, $U_{\text{crit}} = 27$ $p = 16.7 \times 10^{-4}$
7.	Sheltered flutings (Diagn.) (vs.) Unsheltered flutings (Diagn.)	$N_1 = 24$, $N_2 = 12$ $U = 218$, $U_{\text{crit}} = 74$ $p = 7.98 \times 10^{-10}$	$N_1 = 24$, $N_2 = 12$ $U = 218$ $U_{\text{crit}} = 74$ $p = 7.98 \times 10^{-10}$	$N_1 = 24$, $N_2 = 12$ $U = 252$, $U_{\text{crit}} = 74$ $p = 5.92 \times 10^{-5}$
8.	Unsheltered flutings (Diagn.) (vs.) Unsheltered flutings (Ds)	$N_1 = 12$, $N_2 = 12$ $U = 144$, $U_{\text{crit}} = 31$ $p = 3.69 \times 10^{-7}$	$N_1 = 12$, $N_2 = 12$ $U = 144$, $U_{\text{crit}} = 31$ $p = 3.69 \times 10^{-7}$	$N_1 = 12$, $N_2 = 12$ $U = 127$, $U_{\text{crit}} = 31$ $p = 4.28 \times 10^{-4}$
9.	Sheltered reedings (Diagn.) (vs.) Unsheltered flutings (Diagn.)	$N_1 = 6$, $N_2 = 12$ $U = 62$, $U_{\text{crit}} = 9$ $p = 6.7 \times 10^{-3}$	$N_1 = 6$, $N_2 = 12$ $U = 55$, $U_{\text{crit}} = 9$ $p = 4.1 \times 10^{-2}$	$N_1 = 6$, $N_2 = 12$ $U = 54$, $U_{\text{crit}} = 9$ $p = 5.1 \times 10^{-2}$
10.	Unsheltered flutings (Diagn.) (vs.) Unsheltered reedings (Diagn.)	$N_1 = 12$, $N_2 = 6$ $U = 72$, $U_{\text{crit}} = 9$ $p = 5.38 \times 10^{-5}$	$N_1 = 12$, $N_2 = 6$ $U = 72$, $U_{\text{crit}} = 9$ $p = 5.38 \times 10^{-5}$	$N_1 = 12$, $N_2 = 6$ $U = 65$, $U_{\text{crit}} = 9$ $p = 2.3 \times 10^{-3}$

Table 2 (*continued*)

		Conditional thickening	Region growing	Sub-region decomposition
11.	Unsheltered reedings (Diagn.) (vs.) Unsheltered reedings (Ds)	$N_1 = 6,\ N_2 = 6$ $U = 24,\ U_{\text{crit}} = 3$ $p = 0.19$	$N_1 = 6,\ N_2 = 6$ $U = 24,$ $U_{\text{crit}} = 3$ $p = 0.19$	$N_1 = 6,\ N_2 = 6$ $U = 26,\ U_{\text{crit}} = 3$ $p = 0.12$
12.	Unsheltered reedings (Diagn.) (vs.) Unsheltered reedings (WMB)	$N_1 = 6,\ N_2 = 6$ $U = 36,\ U_{\text{crit}} = 3$ $p = 10.8 \times 10^{-4}$	$N_1 = 6,\ N_2 = 6$ $U = 31,$ $U_{\text{crit}} = 3$ $p = 2.05 \times 10^{-2}$	$N_1 = 6,\ N_2 = 6$ $U = 36,\ U_{\text{crit}} = 3$ $p = 10.8 \times 10^{-4}$
13.	Sheltered reedings (Diagn.) (vs.) Unsheltered reedings (Diagn.)	$N_1 = 6,\ N_2 = 6$ $U = 36,\ U_{\text{crit}} = 3$ $p = 10.8 \times 10^{-4}$	$N_1 = 6,\ N_2 = 6$ $U = 36,$ $U_{\text{crit}} = 3$ $p = 10.8 \times 10^{-4}$	$N_1 = 6,\ N_2 = 6$ $U = 36,\ U_{\text{crit}} = 3$ $p = 10.8 \times 10^{-4}$

are smaller in extent than the corresponding areas of the second population. The controversial assumption is stated as alternative hypothesis. To prove or disprove the tested hypotheses we use one-sided statistical test. Thus, whether the *U*-value is much greater than the critical U, the null hypothesis is rejected in favor of the alternative hypothesis.

Through the results reported in Table 2 it is obvious that the cleaning methods attain to eliminate the size of corrosion patterns. This observation is valid for almost all tests, except for the case where unsheltered reedings are cleaned by the DS method. This supports the conclusions derived by the chemical analysis [26,27] according to which DS performs mild cleaning. Therefore, is preferable for the cleaning of unsheltered reedings with flaws and texture irregularities, which demand cleaning methods that minimize material loss. Another objective of this test is to elucidate whether the different conditions of exposure affect the size of the segmented decay areas. According to the results obtained, the black particles detected on sheltered flutings are always larger than the corresponding spots detected on any other of the studied surfaces. In general, decay patterns larger in extent are encountered on sheltered untreated areas. This is expected because crusts of greater thickness prevail there. Comparing the size of decay patterns occurring on reedings and flutings, we can state that decay areas of larger size occur on sheltered flutings. A similar assessment can also be drawn for the unsheltered areas. However, the difference is less significant.

Aiming at investigating whether the results derived by the algorithmic approaches indeed converge with the assessments obtained by chemical analyses we also present the composition results of Table 3. As it can be seen, more severe degradation occurs to untreated surfaces and column flutings. This is reflected on the higher concentration of aluminosilicates and gypsum prevailing on such surfaces. In particular, aluminosilicates contribute significantly to the darkness of degraded areas. This effect further exemplifies the results provided by the *t* tests, which indicate that the depth of deterioration is closely associated with the darkness of corroded areas. Furthermore, the data provided in Table 3 support the conclusions extracted by Tables 1 and 2. More specifically, chemical analyses revealed that untreated areas and sheltered column flutings generally preserve higher concentrations of decay products. Such assessments are also expressed in the statistical tests employed in this work (Tables 1 and 2), which verify that decay areas segmented

Table 3
Results provided by the chemical analysis of the studied surfaces

	Diagnosis	DS	BP	WMB
Sheltered flutings	Gypsum[b], calcite[c], oxalates[e]	Calcite[a], aluminosilicates[f], gypsum[e], oxalates[e]	Calcite[a], oxalates[e], gypsum[e]	Calcite[a], oxalates[e], gypsum[f]
Chemical analyses	Aluminosilicates[d], nitrates[e]			
Sheltered reedings	Calcite[a], gypsum[d], aluminosilicates[d]	Calcite[a], oxalates[e], gypsum[f]	Calcite[a], gypsum[d], aluminosilicates[e], oxalates[e]	
Chemical analyses	Oxalates[e]			
Unsheltered flutings	Calcite[b], aluminosilicates[d], gypsum[d]	Calcite[a], aluminosilicates[d], organic compounds[e]		
Chemical analyses	Oxalates[e], organic compounds[e]			
Unsheltered reedings	Calcite[b], aluminosilicates[e], gypsum[d]	Calcite[a], aluminosilicates[e], organic compounds[e]		Calcite[a], aluminosilicates[e], organic compounds[e]
Chemical analyses	Oxalates[e], barite[e], organic compounds[e]			

[a] >75%.
[b] 50–75%.
[c] 20–50%.
[d] 5–20%.
[e] <1–5%.
[f] <1% (traces).

on surfaces of more severe degradation are characterized by lower distribution of intensity values and larger sizes (due to the larger spatial density of decay patterns induced by their higher concentrations).

5. Remarks and conclusions

In this study, we implemented and tested several image segmentation algorithms with the objective to systematically address the estimation of both the size and topology of degraded areas due to corrosion on stonework. At a subsequent step we also implemented an automated process of defining the Ground Truth matrix and assessing the performance and differences between the segmentation procedures. According to the derived results the Conditional Thickening and the Region Growing Algorithms seem to approach better the detection problem, while the Sub-Region Decomposition and the Adaptive Thresholding schemes follow. A significant result obtained from the performance analysis is that the algorithms' responses differ when processing the digital camera image. In this case, a simple broadband high-pass filtering technique seems to provide quite accurate results. A further significant observation involves the discernibility of the algorithmic approaches to segment decay spots in in-homogenous regions. According to the ROC curves, all detection schemes, except for the Conditional Thickening and the Region Growing, demonstrate a greater efficiency to segment decay spots in rapidly varying backgrounds.

Moreover, the *t* tests and the Mann–Whitney *U* test studied how the cleaning and the structural state are reflected onto the size of decay areas and their relative intensities over the background. These statistical tests revealed that the cleaning methods attain to reduce significantly the size of the segmented decay areas. It is verified that larger in extent decay areas were detected on sheltered surfaces than on unsheltered, as a result of the pollutant accumulation. Regarding the relative intensities of corroded areas over the background, it was revealed that cleaning attained to sufficiently reduce the darkness of the remaining degraded areas. The DS cleaning attains to reduce the darkness of corroded areas, however, they consistently appear darker than decay areas remained after cleaning with the other treatments. Darker decay regions prevail on sheltered untreated flutings, while sheltered untreated reedings and unsheltered untreated flutings follow in severity of degradation. Tests of statistical significance were conducted on decay areas segmented by the three most efficient algorithms. A noteworthy point in this study is the significant convergence between the results of the tested algorithmic schemes. Slight deviations can only be observed in the case of the sub-region decomposition algorithm. These are mainly caused by its tendency to split the segmented regions.

The implementation of machine vision techniques to aid the evaluation of corrosion damage is a challenging issue of particular importance, as it enables non-destructive diagnosis and reduces the diagnosis effect. This work verifies that automated detection schemes contribute to effectively and objectively diagnosing the decay state and reliably evaluating the potential of cleaning approaches. Future research efforts should be focused towards corrosion damage classification according to decay patterns origin and/or type (flaws, material loss, black crusts). Such approaches require thorough analysis of macroscopical images. These classification results along with other criteria posed by the experts can be subsequently used for the development of information systems able to generate integrated descriptions of the studied surfaces and their degradation mechanisms, as well as to support the retrieval of images depicting similar corrosion effects.

Acknowledgements

The authors acknowledge the kind provision of FOM images by Prof. A. Moropoulou and the staff of her Laboratory of Materials Science and Engineering, Technical University of Athens.

References

[1] A. Moropoulou, K. Bisbikou, K. Torfs, R. Van Grieken, F. Zezza, F. Macri, Origin and growth of weathering crusts on ancient marbles in industrial atmosphere, Atmos. Environ. 32 (6) (1998) 967–982.

[2] P. Maravelaki-Kalaitzaki, Black crusts and patinas on Pentelic marble from the Parthenon and Erechtheum (Acropolis, Athens): characterization and origin, Anal. Chim. Acta 532 (2) (2005) 187–198.

[3] P. Maravelaki-Kalaitzaki, D. Anglos, V. Kilikoglou, V. Zafiropulos, Compositional characterization of encrustation on marble with laser induced breakdown spectroscopy, Spectrochim. Acta B 56 (6) (2001) 887–903.

[4] A. Moropoulou, N. Avdelidis, IRT in the investigation of buildings and historic structures, Thermosense XXVI, vol. 5405.

[5] V. Lebrun, E. Bonino, J.F. Nivart, E. Pirard, Development of specific acquisition techniques for field imaging – applications to outcrops and marbles, in: Proceedings of the International Symposium of Geovision on Imaging Applications in Geology, Liege, Belgium, 1999, pp. 165–168.

[6] D. Gelli, D. Virulano, Speed up of shape from shading using graduated non-convexity, in: I. NystrSm, G.S. di Baja, S. Svensson (Eds.), Proceedings of the 11th International Conference, Naples, Italy, 2003.

[7] L. Moltedo, G. Mortelliti, O. Salvetti, D. Vitulano, Computer aided analysis of buildings, J. Cult. Herit. 1 (1) (2000) 59–67.

[8] C. Boukouvalas, F. De Natale, G. De Toni, J. Kittler, R. Marik, M. Mirmehdi, M. Petrou, P. Le Roy, R. Salgari, G. Vernazza, Automatic system for surface inspection and sorting of tiles, J. Mater. Process. Technol. 82 (1–3) (1998) 179–188.

[9] M. Pappas, I. Pitas, Old painting digital color restoration, in: Noblesse Workshop on Non-Linear Model Based Image Analysis, Glasgow, Scotland, 1998, pp. 188–192.

[10] X.E. Gros, J. Bousique, K. Takahashi, NDT data fusion at pixel level, NOT & E Int. 32 (1999) 283–292.

[11] K.Y. Choi, S.S. Kim, Morphological analysis and classification of surface corrosion damage by digital image processing, Corros. Sci. 47 (1) (2005) 1–15.

[12] K.M. Kim, J.J. Park, M.H. Song, I.C.Kim, C.Y. Suen, Design of a binary decision tree for recognition of the defect patterns of cold mill strip using generic algorithm, in: Innovations in Applied Artificial Intelligence: 17th International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems, Ottawa, Canada, 2004, pp. 341–350.

[13] P. Maravelaki-Kalaitzaki, V. Zafiropulos, P. Pouli, D. Anglos, C. Balas, R. Salimbeni, S. Siano, R. Pini, Short free running Nd:YAG laser to clean different encrustations on Pentelic marble: procedure and evaluation of the effects, J. Cult. Herit. 4 (2003) 77–82.

[14] N. Otsu, A threshold selection method from gray-level histograms, IEEE Trans. Syst. Man, Cybern. 9 (1) (1979) 62–66.

[15] J. Dengler, S. Behrens, J.F. Desaga, Segmentation of micro-calcifications in mammograms, IEEE Trans. Med. Imag. 12 (1993) 634–642.

[16] J.C. Russ, The Image Processing Handbook, CRCF Press, Boca Raton, FL, 1992.

[17] A. Hoover, G. Jean-Baptise, X. Jiang, P. Flynn, H. Bunke, D. Goldgof, K. Bowyer, D. Eggert, A. Fit/gibbon, R. Fisher, An experimental comparison of range image segmentation algorithms, IEEE Trans. Pattern Anal. Mach. Intell. 18 (1996) 673–689.

[18] M.N. Gurcan, Y. Yardimci, A.E. Cetin, R. Ansari, Detection of micro-calcifications in mammograms using nonlinear sub-band decomposition and outlier labeling, in: Proceedings of SPIE Visual Communications and Image Processing Conference, San Jose, CA, 1997.

[19] B. Acha, C. Serrano, R.M. Rangayyan, Detection of micro-calcifications in mammograms by seed selection and multi-tolerance region growing, in: Proceedings of the European Medical & Biological Engineering Conference (EMBEC'99), vol. 37, part II, Vienna, Austria, 1999, pp. 984–995.

[20] J.A. Hanley, Receiver operating characteristic (ROC) methodology: state of the art, Crit. Rev. Diagnos. Imag. 29 (3) (1989) 307–335.

[21] C.E. Mertz, Basic principles of ROC analysis, Seminar Nucl. Med. VII (4) (1978) 283–298.

[22] K.H. Zou, WJ. Hall, D.E. Shapiro, Smooth nonparametric receiver operating characteristic ROC curves for continuous diagnostic tests, Stat. Med. 16 (1997) 2143–2156.

[23] L.J. Kitchens, Exploring Statistics: A Modern Introduction, first ed., West Publishing Co., St. Paul, NY, 1987.

[24] J.W. McGree, Introductory Statistics, first ed., West Publishing Co., St. Paul, NY, 1985.

[25] WJ. Conover, Practical Non-Parametric Statistics, second ed., Wiley, New York, 1980.

[26] A. Moropoulou, M. Koui, E.T. Delegou, A. Bakolas, M. Karoglou, E. Rapti, S. Kouris, T. Mauridis, E. Ypsilanti, N.P. Avdelidis, D. Ntae, Conservation interventions planning for the main facade of the National Archaeological Museum, in: Final Research Working Paper, Lab of Materials Science and Engineering, National Technical University of Athens, 2002.

[27] M. Koui, E.T. Delegou, D. Dai, E. Rapti, T. Mavridis, A. Moropoulou, FTIR for the assessment of cleaning interventions on Pentelic marble surfaces, in: Proceedings of the 6th International Symposium on the Conservation of Monuments in the Mediterranean Basin, Lisbon, 2004, CD-Rom Proceedings.