A Practical Use of ROC Analysis to Assess the Performances of Defects Detection Algorithms

Yann Le Meur,[a] Jean-Michel Vignolle,[b] and Jocelyn Chanussot[c]

[a]Trixell, 460 rue du Pommarin, 38430 Moirans, France
Tel. : +33476570081
yann.lemeur@trixell-thalesgroup.com

[b]Trixell, 460 rue du Pommarin, 38430 Moirans, France
Tel. : +33476574292
jean-michel.vignolle@trixell-thalesgroup.com

[c]GIPSA-lab, Departement Image et Signaux, 961 rue de la Houille Blanche, 38402 Saint Martin d'Heres
Tel. : +33476826273
jocelyn.chanussot@lis.inpg.fr

Abstract. Defects detection on images is a current task in quality control and is often integrated in partially or fully automated systems. Assessing the performances of defects detection algorithms is thus of great interest. However, being application and context dependent, it remains a difficult task. This paper describes a methodology to measure the performances of such algorithms on large size images in a semi-automated defect inspection situation. Considering standard problems occurring on real cases, a comparison of typical performance evaluation methods is made. This analysis leads to the construction of a simple and practical ROC-based method. This method extends the pixel-level ROC analysis to an object-based approach by dilating the ground-truth and the set of detected pixels before calculating true positive and false positive rates. These dilations are computed thanks to the *a priori* knowledge of a human defined ground-truth and gives to true positive and false positive rates more consistent values in the semi-automated inspection context. Moreover, dilation process is designed to be automatically suited to the objects shape in order to be applied on all types of defects.

Keywords: target detection, defect detection, detection algorithms performance, ROC curves, object comparison

1 INTRODUCTION

Quality control tasks are some of the main application fields of digital image processing, and particularly detection theory. The amount of new image processing techniques applied to industrial inspection is a relevant proof of the interest taken by both industrials and academics in this problem.

The leading stakes of these techniques are defect detection on textile, wood, or other industrial matters by automated inspection on digital images [1,2]. These images can be acquired by simple optical imaging, X-ray imaging or by non-destructive methods like ultrasounds reflection on the surfaces to be inspected. In this paper, the retrieval of defects on digital X-ray detectors is considered. Digital detectors are now used in X-ray radiography to acquire digital images. The advantages of this fully digital system are obvious: less exposure dose is required than with film systems, the provided images have a better quality, digital format enables an easy storage and transmission, digital processing algorithms can be used in order to enhance diagnostic reliability, etc.

Since the production process of such devices is lengthy and requires human intervention, an important issue is to check the detector's output images quality, and especially to search for potential defects on these images. The detection and localization of such defects can be achieved by image processing algorithms.Defects on digital X-ray detectors produce spurious features on the output images, with various shapes and properties. Their detection thus remains a difficult problem and several algorithms must consequently be considered, evaluated and compared.

Different methods aiming at quantifying detection performances of an algorithm have been described in the literature. In the frame of text detection and recognition, Wolf [3] assesses the performances by rectangle matching and performance graphs. Liu [4] proposes a simple method based on neighborhood inspection to evaluate edge detection performances. Nascimento [5] classifies types of detection errors in order to build a metric for the evaluation of object detection algorithms in surveillance applications. More general methods use common metrics merged in a basic way [6], or thanks to fuzzy logic [7]. Even though they can be useful for specific applications, the reliability of these methods vanishes when considering the task of detecting different objects with various shapes.

This paper proposes a practical view of how defect detection algorithms can be evaluated and gives a response to the assessment of these algorithms based on well-known ROC analysis and object morphology. An overview of the inspection task is followed by a brief description of ROC methodology and an original method derived from this methodology is presented and discussed. Finally, an illustration of how to use such method to process an automated thresholding of detection images is given.

2 THE INSPECTION TASK

Digital X-ray detectors provide large size grayscale images, larger than 3000 by 3000 pixels. The inspection task consists in finding defect on these images. The considered images are acquired with a nominal X-ray dose and without any subject between the X-ray generator and the detector : consequently, images are only composed of the acquisition noise, that we will consider as background, and of potential defects we want to detect. In this case, the goal of detection algorithms is to localize defects areas that a human expert will then check more specifically. The quality control task is not fully automated : the detection algorithms provide a help for a faster and more reliable human decision. This is a typical situation for an industrial context : the detection algorithms are designed to catch the attention of a human operator and only focus on potential defects areas in order to make the inspection task less tedious. The evaluation of the algorithms performances must take this context into account and provide a measure suitable for various defects. The main consequences introduced by this specific application are :

- the perfect location of all the defective pixels is not required : actually, the human expert just needs some pixels at each defect location to identify it. On the other hand, the whole set of defects must be identified by the detection of at least one pixel lying inside each target.
- the borders of each defects are approximately defined : since the design of a ground-truth remains subjective, the areas near the borders of a defect can be seen either as defective pixels or as background pixels. Moreover, some defects have naturally fuzzy borders ; a precise and certain ground truth can thus not be defined.

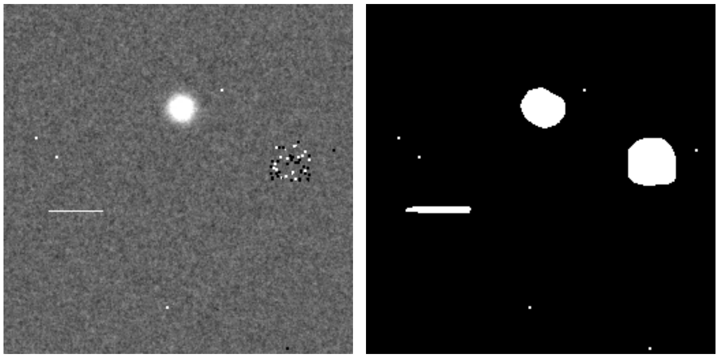
These observations implies particular definitions and use of ROC methodology that we will describe in section 4.

In our specific application, the defects to detect are of two kinds :

- punctual defects : isolated pixels or little clusters of pixels with abnormal statistics (strong luminance)
- extended defects : spatially-correlated set of pixels that are not statistically atypical when considered individually. They come in the shape of lines, columns, spots or gathered clusters.

Fig 1 shows examples of synthetic defects with the associated defect map designed by a human expert. These examples have been ”hand-made” to illustrate extreme cases of defects that could occur in any kind of imaging system.

This figure spotlights the various ways to build the defect maps, depending on the type of defect. For punctual defects, the defect map is defined precisely, whereas for extended defects, like the fuzzy spot, the defect map is more subjective. When several clusters are gathered (as on the right side of fig 1(a)), the defect map includes the clusters and some non-defective pixels in one single object. It also underlines the scale adaptation required to identify high-level structures of defects. As a matter of fact, the clusters of defective pixels on the right side of fig 1(a)) are not identified as several punctual defects, but rather as a single defective area made of punctual defective pixels. These remarks should be considered when designing a method to assess the performances of defect detection algorithms in such semi-automated inspection task.



(a) Various defects (b) Defect map

Fig. 1: Different kinds of defects and corresponding defect map

		Detection System *Defect detected*	 *No Defect detected*	
Ground Truth	*Defect*	True positive TP	False negative FN	*TP+FN* = *Total positives*
	No Defect	False positive FP	True negative TN	*FP+TN* = *Total negatives*

Fig. 2: The confusion matrix represents the true positive and the false positive for a defect detection task

3 THE ROC ANALYSIS

3.1 Definitions and ROC curves

Introduced in the early 80's, the ROC (Receiver Operating Characteristic) methodology has become a standard technique to evaluate detection performances. It was firstly used to measure diagnostic performances of medical imaging systems, especially in radiologic imaging [8–10]. It has since been extended to various detection systems.

For a single target (a defective area in our case) problem, the ROC analysis consists in measuring the binary response of the detection system (target present or not) to one stimulus, in our case an image, by calculating the true positive rate tpr and the false positive rate fpr with :

$$tpr = \frac{\text{true positive}}{\text{total positives}} \quad (1)$$

$$fpr = \frac{\text{false positive}}{\text{total negatives}} \quad (2)$$

Fig 2 presents the classical representation of a confusion matrix.

A couple $(fpr; tpr)$ corresponds to one point in the ROC plane. ROC curves are plot by changing the parameters of the detection systems and computing tpr and fpr at each value of the parameters set. The ROC analysis is an appropriate tool to deal with detection performances since it takes the prevalence of each class into account and provides two antagonist intuitive measures that are meaningful for the system calibration.

In a defect detection context, there can be many defects on one image. The true positive and false positive can be computed on this image following the Free-ROC methodology [11] (FROC). Free-ROC is the extension

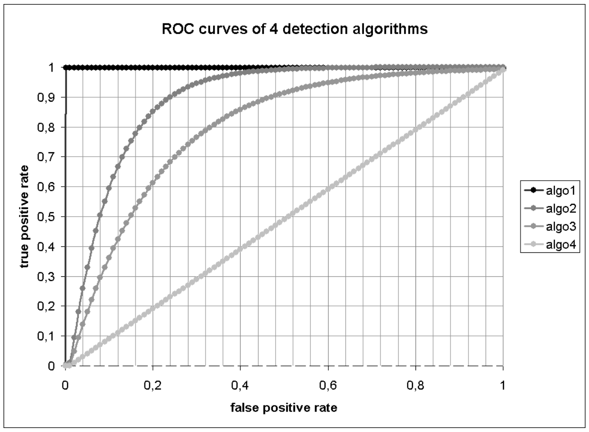


Fig. 3: Examples of ROC curves

of ROC methodology to target localization, while ROC only deals with target detection. In our application, the advanced theory of Free-ROC is not needed so we will only consider basic tools of ROC methodology. Considering an image with several targets - the defects - and a defect detection algorithm providing a pixel-by-pixel classification with two class (defect or no defect), there are four cases for each pixel p_i of the image :

- p_i is classified as defect and is a defect in the ground truth image: it is a true positive also called hit, or recall
- p_i is classified as background (no defect) and is a background pixel in the ground truth image: it is a true negative
- p_i is classified as defect and is a background pixel in the ground truth image: it is a false positive also called false alarm
- p_i is classified as background and is a defect pixel in the ground truth image: it is a false negative

The main advantage of ROC analysis is that the two quantities, tpr and fpr are normalized to the number of positive and negative samples, respectively. Then, unlike the traditional measures like accuracy (the percentage of pixels correctly classified), tpr and fpr cannot be biased by a small prevalence of one class compared to the other.

To spotlight this statement, let us consider a detection task on a 1000×1000 pixels image with one single defective pixel. An algorithm that systematically detects nothing is actually very accurate : 99.9999% of the pixels are correctly classified. On the other hand, both its fpr and its tpr will be 0%, thus revealing really bad detection performances. As a conclusion, the sole accuracy of the algorithms does not provide enough information to ensure a reliable estimation and is biased in this case by the small prevalence of the defect class.

ROC analysis features in one curve the sensibility tpr of the detection system *versus* fpr ($fpr = 1 - specificity$), which are the two quantities of interest in a control quality context : it indicates how many false alarms are generated by the system for a given detection sensibility. Moreover, a ROC curve provides the dynamic behavior of the system with respect to a change of the decision threshold. This information can be used to choose between two detection systems : the ROC curve of the best detection system is the one which is always over the other curve in the ROC plane.

The ROC curves of four detection algorithms are displayed on fig 3. The perfect detection algorithm is ”algo1” : its ROC curve is a step function (100% of tpr for any fpr). In the case where the two classes, defects and background, are equally distributed, an algorithm corresponding to a random decision has the ROC curve ”algo4” (the ascending diagonal of the ROC plane). Between random and ideal decision, ”algo2” performs better than ”algo3”.

A practical measure of the global performance of an algorithm is given by the area under its ROC curve. This Area Under Curve (AUC) is commonly used to quantify with one single number the overall performance of a detection algorithm [8, 12, 13].

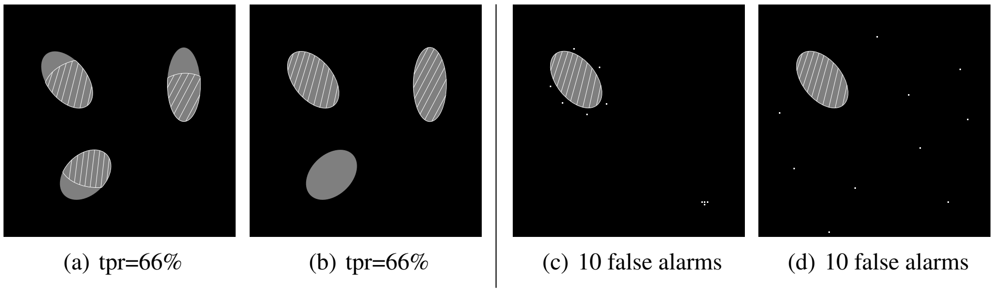

(a) tpr=66% (b) tpr=66% (c) 10 false alarms (d) 10 false alarms

Fig. 4: tpr and fpr meaning : on each image the targets to detect are in gray and the detected pixels in white. The first two images have the same tpr and the last two have the same number of false alarm pixels

4 THE COMPARISON OF MASKS

In section 3, the ROC analysis was presented as a useful tool to assess the performances of a detection algorithm. However, a major problem remains : how can tpr and fpr be estimated ? In other words: which pixels should be considered as true positive or false positive ?

At each decision threshold, tpr and fpr are calculated by comparing a binary detection mask (with ones for defects and zeros for the background) with a ground-truth mask. In the following, the detected defect mask is called the test mask, $M_{i,j}$. It results from a pixel-wise decision produced by the detection algorithm. The manually-designed ground-truth mask is called the target mask, $T_{i,j}$. In practice, the simplest way to compare these two masks is to make a pixel-level comparison, thus exactly fitting the definitions of false positive and true positive given in section 3.

As discussed in section 2, the human expert does not need the detector to provide a $+/-1$ pixel precision for the localization of the defects. Based on this assumption, alternative methods have been proposed to calculate tpr and fpr : Theiler [14] suggests to transform the test mask in order to consider each object of the target mask as 100% detected as soon as at least one pixel is detected in this object. In a similar way, Harvey [15] proposes to dilate the test mask by a fixed factor, in order to include some of the "near hit" pixels of the test mask in the tpr count.

The built of such alternative methods take significance in our semi-automated inspection context where we have to be more specific on the definition of a true positive and a false negative. Firstly, considering the true positives: on a multi-target situation, tpr should be of 100% if the detected pixels leads to the visual localization of all targets, even if all the defective pixels are not detected. On the example of fig 4(a) about 66% of the defective pixels are detected (hatching areas) but the detection allows the expert to localize all the targets. The global tpr on fig 4(b) is also of 66%, but for the human expert the detection is clearly worse since the bottom target has been completly missed. Now considering false alarms, we should make distinction between isolated false alarms, false alarms close to a target ("near hit") and clusters of false alarms. For the human expert, the number of false alarms is the number of observation windows, called AOI ("Area of Interest") that the system will display to be checked. Each window displayed with no true defect will be considered as false alarm. On the example of fig 4, the two images of fig 4(c) and 4(d) have both ten false alarms pixels. But for the fig 4(c), the human expert will consider the detection to have only one false alarm : the few detected pixels near the target cannot be seen as false alarm because the associated AOI will include a part of the target. The other detected pixels forms a cluster that will be embedded in only one AOI that will give to the only false alarm of the image. On fig 4(d), there are only isolated false alarms : the expert has to check ten AOIs which do not include a defect: the detection system then raises ten false alarms.

Taking into account these observations, we propose in this paper a method to compare binary masks that brings up less penalization to detected pixels near the objects borders. These pixels are considered as false alarms by all the previous methods. Moreover, our method does not require any parameter to be tuned and can be applied to the cases where multi-targets of different sizes and shapes are to be found in a single image.

In the following parts, three masks comparison methods are described and discussed, namely the simple pixel level comparison, Theiler's method and our proposed algorithm, focusing on the problems raised by the

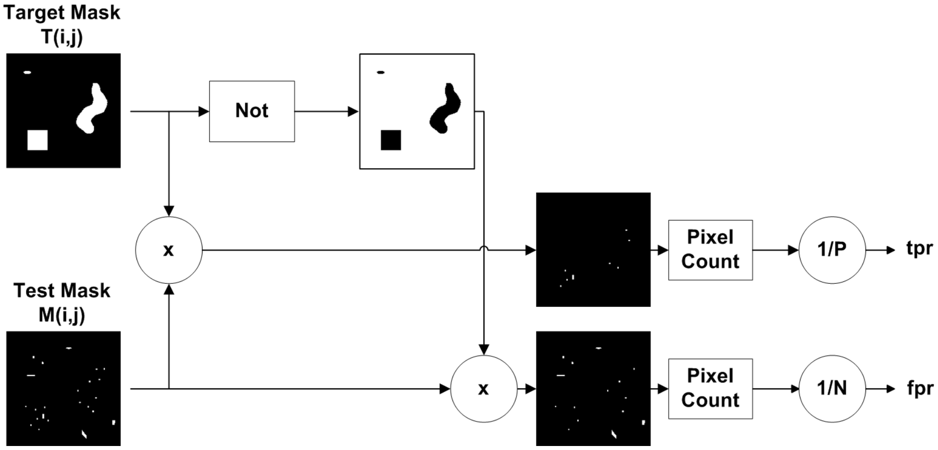


Fig. 5: Pixel level masks comparison:computes tpr and fpr by direct comparison between target and test masks. ”Not” stands for binary complement and ”Count” returns the number of white pixels in the input image

Defect	fpr (%)	tpr (%)
Punctual	0.12	50
Spot	0.03	3.8
Cluster	0.12	1.5

Table 1: tpr and fpr computed with the pixel level masks comparison

semi-automated inspection task. Harvey’s comparison method requires a strong *a priori* knowledge of the size of the targets. It is thus not further developed in this paper.

4.1 Pixel level masks comparison

The first and most intuitive method is to compute the binary comparison between target and test masks, without any preprocessing. Considering the binary target mask $T_{i,j}$ with P defective pixels (pixels with value 1) and N background pixels (with value 0), the pixel-level mask comparison is described by fig 5. The ”Pixel Count” box returns the number of pixels with value 1 in the input image, and the ”Not” box stands for binary complement operator. tpr and fpr are thus computed as follows :

$$tpr = \frac{1}{P}\sum_{i,j} M_{i,j} \cdot T_{i,j} \tag{3}$$

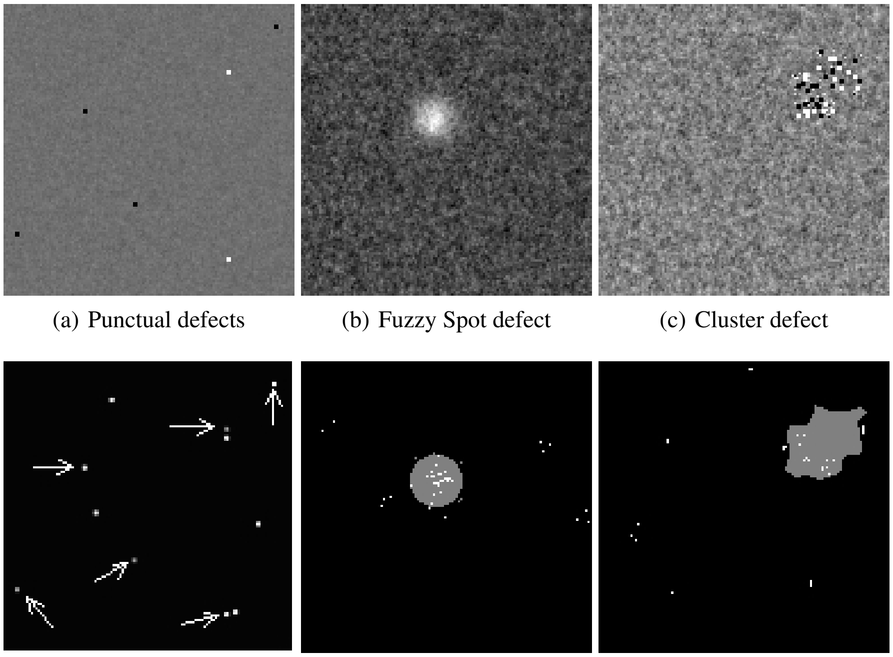
$$fpr = \frac{1}{N}\sum_{i,j} M_{i,j} \cdot (1 - T_{i,j}) \tag{4}$$

The pixel level masks comparison is computed on the three synthetic defects of figs 6(a), 6(b) and 6(c). The corresponding ground-truth is displayed in gray on the second line of this figure, with the test mask appearing in white. Missed pixels on these latter images are in gray and detected pixels are always in white, no matter if they are false alarms or true positives. The three defects chosen are :

- Punctual defects : six isolated defective pixels
- Fuzzy spot defect : a bright spot with fuzzy borders
- Cluster defects : a cluster of defective pixels forming an area identified by the human expert as one single defective object

For clarity purpose, the ground-truth pixels for the punctual defects are pointed on fig 6(d) by arrows.

Tab 1 presents the computed tpr and fpr for the three defects with this pixel level masks comparison.



(a) Punctual defects (b) Fuzzy Spot defect (c) Cluster defect

(d) Punctual test (e) Spot test (f) Cluster test

Fig. 6: Three kinds of defects and corresponding target and test masks

For the isolated pixels, the simple pixel level masks comparison provides satisfactory results : for these kind of defects, the exact location is required and false alarms are raised even if the detection is close to the defect. In the practical case, the system must be very precise for this kind of defects in order to catch the human expert's attention on the right pixel because of the very small size of the defects. In this case the pixel level comparison performs well, giving a tpr of one out of two and a fpr which correctly represents the number of times the expert will focus on a non interesting pixel.

The situation of the fuzzy spot defect is more ambiguous. In the presented example, some pixels inside the target are actually detected, but not all of them. In the mean time, the ground-truth is set as a circular area which includes the fuzzy spot. In this case the algorithms's detection is almost perfect : the amount of good detections and their location at the center of the defect are sufficient information for the human expert to properly identify the defect. But due to the ground-truth definition, the fpr and tpr computed at a pixel level are far from the fairly good expected values. Moreover, one can consider the detected pixel at the bottom left of the defect as a near hit and then should not be treated as a false alarm. As a matter of fact, the assessment of detection performances faces the high subjectivity linked to the design of the ground-truth, especially in this case where the defect's borders are fuzzy. This subjectivity is not integrated in the pixel level comparison.

Similar remarks hold for the example of cluster defect. Again, some pixels (left and right sides of the cluster) are counted as false alarms while they are too close to the borders to be actually objectively considered as such. Secondly, many pixels are detected inside the cluster (mainly the punctual defects inside this cluster). Nevertheless the tpr hardly reaches 2% whereas, thanks to these detected pixels, a human expert would identify the whole defect area.

These examples underline the unsuitability of the direct pixel level comparison to assess detection performances in a complex context, mainly because of the two reasons which were underlined in section 2.

4.2 Theiler's mask comparison

As the simple pixel level comparison leads to inappropriate assessment of detection performances, there is a need to derive a technique which somehow mimics the human expert. In this way, Theiler proposed a metric to perform higher level interpretation of the test mask. This techniques lies on a ”filling-in” process : all

Defect	fpr (%)	tpr (%)
Punctual	0.12	50
Spot	0.03	100
Cluster	0.12	100

Table 2: tpr and fpr for Theiler's masks comparison

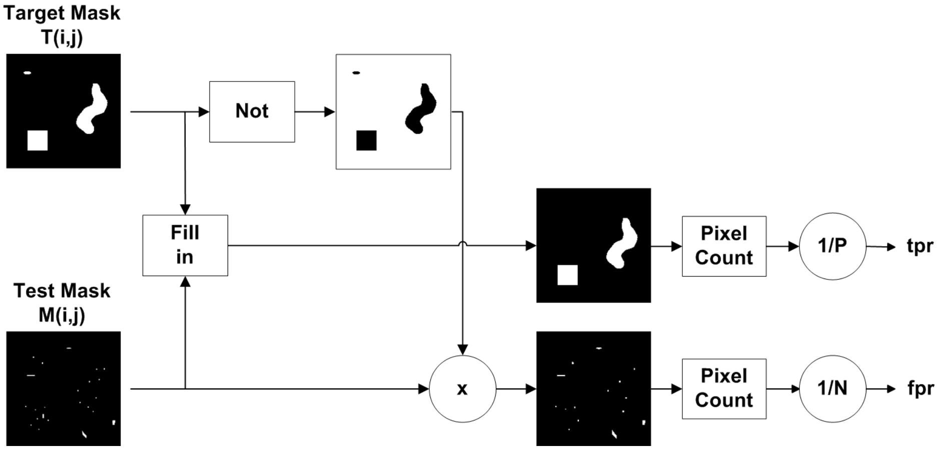


Fig. 7: Theiler's masks comparison

the pixels of a target are considered as detected if at least one of them is actually detected by the detection algorithm (see fig 7). tpr and fpr are thus computed as follows :

$$tpr = \frac{1}{P}\sum_{i,j} Fill(M_{i,j}) \cdot T_{i,j} \tag{5}$$

$$fpr = \frac{1}{N}\sum_{i,j} M_{i,j} \cdot (1 - T_{i,j}) \tag{6}$$

where $Fill$ is the filling-in operator.

Following this approach, tpr and fpr are computed on the defects of fig 6. Corresponding results are displayed on tab 2.

For punctual defects, Theiler's method provides satisfactory results, similar to those obtained with the pixel level mask comparison.

For the two other examples, the fpr according to Theiler's comparison is unchanged but the tpr is now of 100%. As a matter of fact, in each case, at least one pixel is detected inside the target. This strategy is well suited in some cases : for the fuzzy spot defect, the detected pixels are centered on the defect and are ditributed over an area which is not too different from the true defect. Then, the expert will take all these detected pixels as one detection which permits to find the defect : the tpr of 100% is a correct measure of detection performances.

On the other hand, the limits of this filling-in strategy arise when considering the cluster defect. In this case, mainly the bottom part of the defect is detected while the top part has no detection. This kind of test mask opens up to the expert the possibility to miss one part of the defect due to a lack of detected pixels on the whole area of the ground-truth. This problem may occur especially on extended defects and in a multi-target context.

For the fpr calculation, it follows the same rule as the pixel level comparison, with the drawbacks described in previous section.

To conclude, Theiler's comparison extends the pixel level approach to an object approach by simply adding to the test and target mask comparison a filling-in process. On single target context, and with small targets compared to the image size, this comparison performs an acceptable assessment in the inspection context.

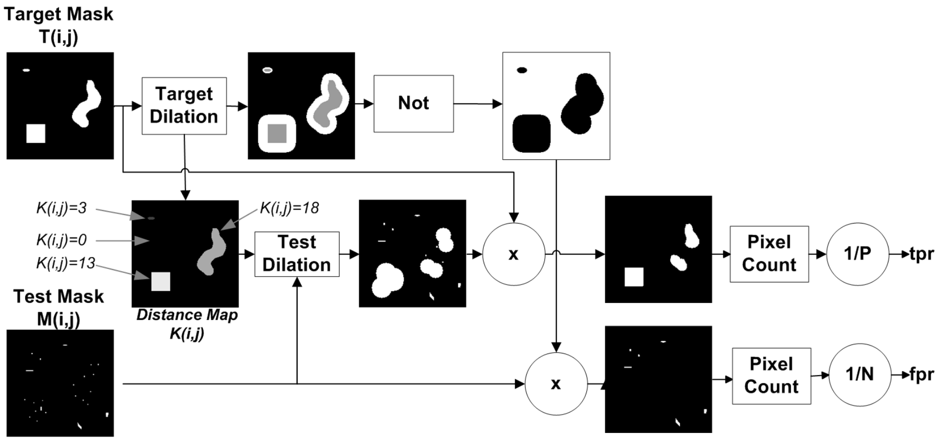


Fig. 8: Original soft masks comparison

Nevertheless, it does not take the distribution of the detected pixels inside the target into account, which is a criterion to look after before declaring the target as 100% detected. Moreover, the false alarms computation remains inappropriate.

4.3 The proposed soft mask comparison

In this section, we present an original method which provides a practical response to the performances assessment requirements and overcomes the limitations of the standard methods previously described. The principle of our method is illustrated by fig 8.

The method requires two new operators : target dilation and test dilation.

- *Target dilation*: the euclidean distance transform of the target mask T is computed [16]. Then the maximum distance d_k for each target α_k is determined thanks to the distance transform computed before. This maximum corresponds to the directed Hausdorff distance [17] between the target and the background. The value of Hausdorff distance is stored in a distance map $K_{i,j}$. Let $p_{i,j}$ a pixel of $K_{i,j}$:

$$K_{i,j} = \begin{cases} d_k & \text{if } p_{i,j} \in \alpha_k \\ 0 & \text{otherwise} \end{cases}$$

  Each object of the target mask is then dilated by a circular structuring element of radius $K_{i,j}$.

- *Test dilation*: each detected pixels $p_{i,j} \in M_{i,j}$ is dilated by a circular structuring element of radius $K_{i,j}$.

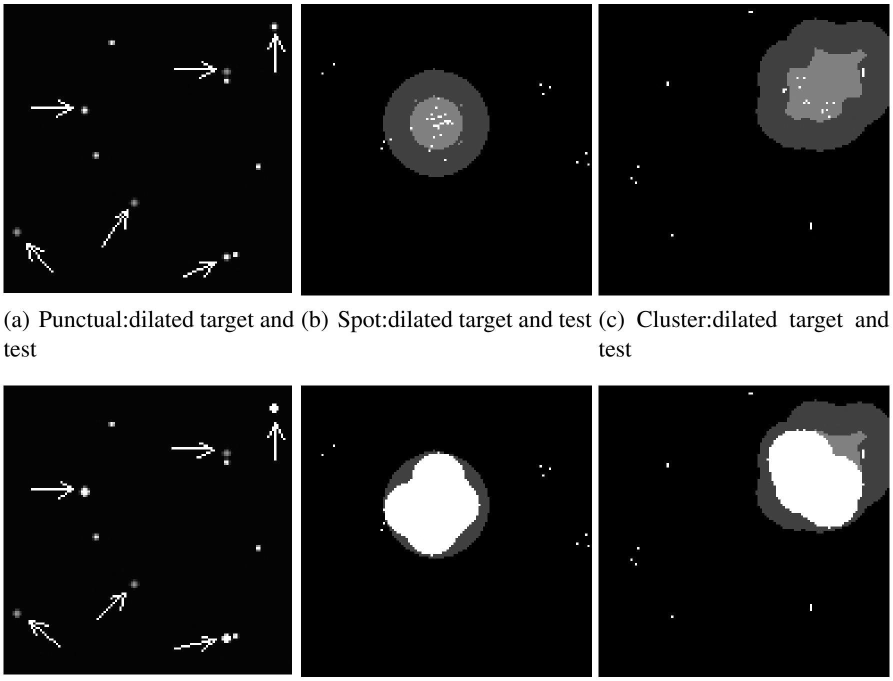
The tpr and fpr are thus computed as follow :

$$tpr = \frac{1}{P} \sum_{i,j} TestDil(M_{i,j}, K_{i,j}) \cdot T_{i,j} \quad (7)$$

$$fpr = \frac{1}{N} \sum_{i,j} M_{i,j} \cdot (1 - TargDil(T_{i,j})) \quad (8)$$

where $TestDil$ stands for the test dilation process and $TargDil$ for the target dilation process.

The results of the proposed soft comparison method on the three test defects are shown on fig 9. The first line of this figure represents the target dilation required for the computation of fpr: the initial target is in light gray, and the dilated target in dark gray while the detected pixels are in white. The target dilation is designed to expand the target's borders, preserving the global shape of the target. Here, the dilation technique leads to approximatively double the distance between the Hausdorff distance of each target. Two questions may then be raised :



(a) Punctual:dilated target and test (b) Spot:dilated target and test (c) Cluster:dilated target and test

(d) Punctual:dilated target and dilated test (e) Spot:dilated target and dilated test (f) Cluster:dilated target and dilated test

Fig. 9: Dilated target mask superposed with test mask (first line) and dilated target superposed with dilated test mask (second line)

1. Why shall the dilation factor depend on the target's size?
2. Why shall the dilation factor be set as described (i.e. one time the maximum distance to the borders of each target)?

The following answers can be given:

1. A large defects is observed at a larger scale. Consequently, the ground-truth is less precise than for a very small defect. As a conclusion, a larger error for the near hit pixels (not counted as false alarms) should be allowed.
2. The chosen distance is the simplest and the least arbitrary size that can be determined in order to approximately double the size of the target. Thus, the target dilation process is an auto-adapted scheme which does not require any parameter.

The result of this target dilation is that some detected pixels near the cluster or spot defects are not considered as false alarms any longer since they now lie within the dilated target mask. The second line of fig 9 shows the result of the test dilation on detected pixels (in white), with the dilated target mask superposed in gray level. Each detected pixel lying in an initial target (in light grey) is dilated before the computation of tpr. This dilation mimics the human expert's behaviour : the focus is not only on the detected pixels but also on the surrounding pixels. Then, it is consistent to consider the area around these detected pixels for the computation of tpr. A consequence of this test dilation is that only some central points have to be detected in order to reach 100% tpr for a target. Typically, the detection of the skeleton [18, 19] of the target is sufficient to reach such tpr, which is in accordance with visual inspection task : the central points of a target are sufficient to indentify the full target.

The corresponding values of fpr and tpr are shown on tab 3. For the punctual defects, there is no change compared to the previous methods. For the spot defect, tpr reaches 100%, which is in great accordance with

Defect	fpr (%)	tpr (%)
Punctual	0.12	50
Spot	0.03	100
Cluster	0.07	80

Table 3: tpr and fpr for Soft masks comparison

a human interpretation. fpr has slightly dropped due to the pixels at the bottom of the defect which are excluded from fpr : indeed they are ”near hit”. For the cluster defect, fpr has dropped for the same reason, while tpr now reaches 80% : as a matter of fact, the top part of the defect is not considered as detected by our method. This is a relevant interpretation since the cluster defects is extended and is made of two defective areas gathered where only the bottom part is actually detected by the algorithm.

In the presented cases, the proposed soft mask comparison gives tpr and fpr results which are consistent with the human expert wishes for detection performances assessment. This method performs a morphological dilation of the target and test masks and is auto-adapted to the multi-target problem since the dilation factors are computed only one time for each target without any parameter.

5 USE OF PROPOSED METHOD FOR AUTOMATED THRESHOLDING

Our proposed soft mask comparison method to compute tpr and fpr has been introduced in the previous section. In this section we will show how to use this method to make an automatic thresholding of images. Detection algorithms usually provide grayscale images where bright pixels are defective pixels and dark ones are background pixels. To get a test mask from this grayscale detection image, we need to set a decision threshold in order to binarize the detection image.

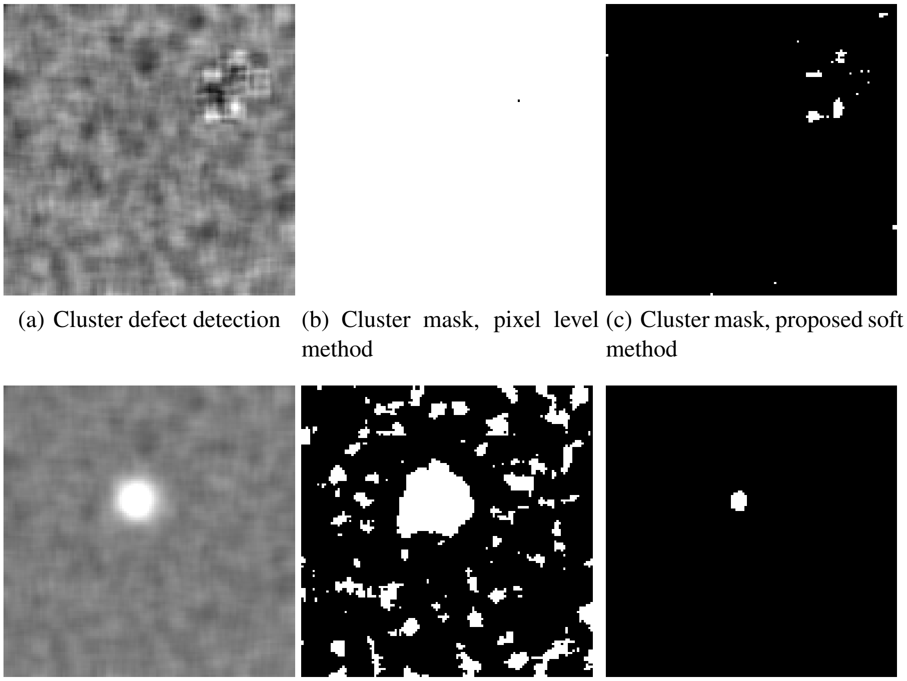
When the target mask is known we can use the ROC methodology to automatically set the decision threshold at a value leading to a given tpr or fpr. In this situation, the soft mask comparison allows to get test masks which are more consistent with the given tpr and fpr.

Considering the defects introduced on fig 6, we need the thresholded image of a grayscale image provided by a detection algorithm. We want the thresholding image at a tpr of 100% (full target image) with the minimum value of fpr. This is a typical image observed to evaluate the detection performances of an algorithm. Fig 10 shows the full target images obtained, for cluster and gauss defect, with two different mask comparison methods introduced earlier : the pixel level method (see 4.1) and the proposed soft method (see 4.3).

For the cluster defect, fig 10(b) spotlights the high sensibility of the pixel level method : to get a $tpr = 100\%$ with our detection algorithms (images of the left column of fig 10) , nearly all the pixels of the image should be detected. Consequently, numerous false alarms are raised, and one could believe that the algoritm performs pretty bad on this defect. It is not the case, and the soft mask comparison (fig 10(c)) allows to avoid such mistake in this case : the thresholded image shows perfect detection in our semi-automated inspection context (the detected pixels are sufficient to localize the whole defect) with only a few false alarms. The same conclusion can be made with thresholded images of fig 10(e) and 10(f). In these two cases, the pixel level comparison, due to its pixel sensibility, leads to overdetection (too many false alarms) while the proposed method, using the new definition of tpr, gives relevant binarized images.

6 POSSIBLE EXTENSIONS

Our method is a first step towards a high-level mask comparison which will be fully adapted to the post-processing inspection made by the human expert. Some immediate extensions of this method may be developed. Firstly, for the fpr computation, we should take into account the size of the AOIs that will be presented to the expert for a visual inspection. Our method makes a pixel by pixel count of false alarms, which corresponds to a pixel-by-pixel inspection of these alarms, i.e. a size of 1 pixel for the AOI. We should rather consider the real size of the inspection system to gather clusters of false alarms in one single false alarm. This could be made by dividing the image in square windows of the same size than the AOIs and by counting the number of windows where false alarms actually occur. The number of false alarms would then measure the number of times an AOI without any real defect has nevertheless to be checked. This is a more meaningful way to assess fpr in our context.



(a) Cluster defect detection (b) Cluster mask, pixel level method (c) Cluster mask, proposed soft method

(d) Gauss defect detection (e) Gauss mask, pixel level method (f) Gauss mask, proposed soft method

Fig. 10: Thresholded images at $tpr = 100\%$ with respect to two mask comparison methods : pixel level and proposed soft mask comparison

Secondly, the tpr has been normalized by the number P (see eq 7) of defective pixels in the image. This choice has been made in order to give a quick interpretation of tpr, but it can be too restrictive in certain cases. Considering an image made of two defects : one of large size and one punctual (one single defective pixel), respectively. A detection algorithm that correctly detects the first defect, but misses the second one, will achieve a fairly high tpr whereas one target out of two has actually been missed. To avoid such situations, we should rather make an object-based normalization : the tpr is computed for each target and the overall tpr for the image is computed by averaging all these rates. Then targets of different sizes with the same share of well detected pixels would then have the same impact on overall tpr value. Moreover, if the AOI size is known, we should rather define a constant dilation factor for the test dilation which fits this size.

To conclude, several improvements of the proposed method can be made by using *a priori* knowledge potentially available for the different stages of the detection system. However, the global performances assessment scheme remains unchanged since we yet consider fuzzy areas for fpr and extended detected areas to compute tpr.

7 CONCLUSION

The inspection of defects on large size images is a very fastidious task. Thus, in order to help the human expert, many automated processes and image processing algorithms have been developed to detect potentially defective areas. Assessing the actual quality and performances of these detection algorithms is then of the utmost importance and must be dealt with respect to the inspection context. ROC-analysis is a proven methodology to compare such algorithms, but it has some limitations when facing complex situations (various sizes/shapes/types of defects). To overcome these limitations, we propose a method to compute true positive and false positive rates, in a way that is consistent with semi-automated inspection application. This method uses simple object-based morphological dilations to extend the pixel-level definitions of ROC quantities to more object-related ones. Thus, fuzzy-areas are automatically defined around each object to exclude near-hit from false alarms count. In the mean time, true positives are linked to the visual inspection problem by defining dilation scheme to mimic human expert inspection. This way to use ROC methodology on practical cases provides more reliable assessment of defect detection algorithms and then allows a better calibration of semi-automated quality control systems.

References

[1] A. Kumar and G. K. Pang, “Defect detection in textured materials using optimized filters,” *IEEE Transactions on Systems, Man, and Cybernetics - Part B : Cybernetics* **32**(5), 553–570 (2002).

[2] D.-M. Tsai and T.-Y. Huang, “Automatic surface inspection for statistical textures,” *Image and Vision Computing* **21**, 307–323 (2003).

[3] C. Wolf and J.-M. Jolion, “Object count/area graphs for the evaluation of object detection and segmentation algorithms,” *Int. J. Doc. Anal. Recognit.* **8**(4), 280–296 (2006).

[4] G. Liu and R. Haralick, “Assignment problem in edge detection performance evaluation,” in *Proceedings of Conference on Computer Vision and Pattern Recognition CVPR 2000*, 1026–1031, IEEE Computer Society, (Hilton Head, SC, USA) (2000).

[5] J. C. Nascimento and J. S. Marques, “Novel metrics for performance evaluation of object detection algorithms,” in *1st ISR Workshop on Systems, Decision and Control Robotic Monitoring and Surveillance, Lisbon*, (2005).

[6] V. Y. Mariano, J. Min, J.-H. Park, R. Kasturi, D. Mihalcik, H. Li, D. Doermann, and T. Drayer, “Performance evaluation of object detection algorithms,” *International Conference on Pattern Recognition* **03**, 30965 (2002).

[7] J. M. James Keller and P. Gader, “A fuzzy logic approach to detector scoring,” in *Fuzzy Information Processing Society - NAFIPS, 1998 Conference of the North American*, **20**, 339–344 (20-21 August 1998).

[8] J. Hanley and B. McNeil, “The meaning and use of the area under a receiver operating characteristic (ROC) curve,” *Radiology* **143**(29), 29–36 (1982).

[9] C. E. Metz, “ROC methodology in radiologic imaging,” *Investigative Radiology* **21**, 720–733 (1986).

[10] T. Fawcett, “ROC graphs: Notes and practical considerations for researchers,” tech. rep., HP Labs (2004). http://home.comcast.net/∼tom.fawcett/public_html/papers/ROC101.pdf.

[11] J. M. Irvine, “Assessing target search performance: the free-response operator characteristic model,” *SPIE Optical Engineering* **43**, 2926–2934 (2004).

[12] P. A. Flach, “Tutorial on the many faces of roc analysis in machine learning,” in *International Conference of Machine Learning*, (http://www.cs.bris.ac.uk/ flach/ICML04tutorial/) (4 Juillet 2004).

[13] C. Cortes and M. Mohri, “AUC optimization vs. error rate minimization,” in *Advances in Neural Information Processing Systems (NIPS 2003)*, **16**, MIT Press (2003).

[14] N. A. D. James Theiler, Neal Harvey and J. M. Irvine, “Approach to target detection based on relevant metric for scoring performance,” *33rd Applied Imagery Pattern Recognition Workshop (AIPR’04)* **00**, 184–189 (2004).

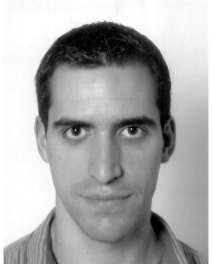
[15] N. R. Harvey and J. Theiler, “Focus-of-attention strategies for finding discrete objects in multispectral imagery,” in *Proceedings of the SPIE, Volume 5546*, S. S. Shen and P. E. Lewis, Eds., **5546**, 179–189 (2004).

[16] A. Jain, *Fundamentals of Digital Image Processing*, Prentice Hall (1995).

[17] W. Rucklidge, *Efficient Visual Recognition Using the Hausdorff Distance*, Springer-Verlag New York, Inc., Secaucus, NJ, USA (1996).

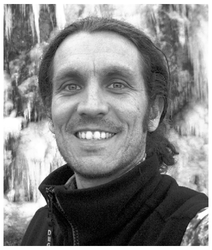
[18] H. Blum, “A Transformation for Extracting New Descriptors of Shape,” in *Models for the Perception of Speech and Visual Form*, W. Wathen-Dunn, Ed., 362–380, MIT Press, Cambridge (1967).

[19] R. C. Gonzales and R. E. Woods, *Digital Image Processing Second Edition*, Prentice Hall (2002). ISBN 0201180758.



Yann Le Meur graduated in electrical engineering from the Grenoble Institute of Technology (INP Grenoble), France, in 2004 and received the same year his M.Sc degree in Signal and Image Processing from INP Grenoble. In 2004, he leaded a six month M.Sc thesis project at Centre National d'Etudes Spatiales (french space agency), Toulouse, France where he worked on multi-temporal remote sensing images. He is now a PhD candidate at GIPSA-Lab (Grenoble Image Speech Signals and Automatics laboratory) and Trixell, Moirans, France. His research interests include image processing, objects detection, image statistical analysis, image quality assessment and data fusion, especially kernel based methods.

Jean-Michel Vignolle graduated in general engineering from the Ecole Centrale Paris, France, in 1987, with a speciality in bio-engineering, and received the same year his M.Sc. degree in Spectrochemical analysis methods, from Paris VI University. In 1988-89 he worked in THOMSON Central Research Labs on materials for Neural Networks Hardware implementation, then fiber optics sensors. In 1990 he joined THALES AVIONICS and was responsible for various LCD display design for projection en direct view. His activities included microelectronics design, electrical design, mechanical design and optical design. In 1998 he has joined TRIXELL as technical project manager on various X-Ray detector design projects. Since 2002 he has been in charge of the image group, a group of engineers dedicated to the development of image processing, image correction algorithms, image quality measurement tools and methods.



Jocelyn Chanussot graduated in electrical engineering from the Grenoble Institute of Technology (INP Grenoble), France, in 1995. He received the Ph.D. degree from the Savoie University, Annecy, France, in 1998. He was with the Automatics and Industrial Micro-Computer Science Laboratory (LAMII). In 1999, he worked at the Geography Imagery Perception laboratory (GIP) for the Délégation Générale de l'Armement (DGA - French National Defense Department). Since 1999, he has been with INP Grenoble as an Assistant Professor (1999-2005), an Associate Professor (2005-2007) and a Professor (2007-) of signal and image processing. He is conducting his research at GIPSA-Lab (Grenoble Images Speech Signals and Automatics laboratory). His research interests include statistical modeling, classification, image processing, nonlinear filtering, remote sensing and data fusion. Prof. Chanussot is currently serving as an Associate Editor for the IEEE Transactions on Geoscience and Remote Sensing and for Pattern Recognition. He is the Co-chair of the GRS Data fusion Technical Committee and a Member of the Machine Learning for Signal Processing Technical Committee of the IEEE Signal Processing Society. He has authored or co-authored over 65 publications in international journals and conferences. He is a Senior Member of the IEEE.

List of Figures