Evaluation criterion of thermal light ghost imaging based on the receiver operating characteristic analysis

WEN-KAI YU,1,2 XU-RI YAO,1,2 XUE-FENG LIU,1 LONG-ZHEN LI,1,2 AND GUANG-JIE ZHAI1,*

1Key Laboratory of Electronics and Information Technology for Space System, Center for Space Science and Applied Research, Chinese Academy of Sciences, Beijing 100190, China
2University of Chinese Academy of Sciences, Beijing 100049, China
*Corresponding author: gjzhai@nssc.ac.cn

Received 24 December 2014; revised 13 April 2015; accepted 13 April 2015; posted 14 April 2015 (Doc. ID 231405); published 8 May 2015

The performances of different thermal ghost imaging (GI) algorithms are compared in an experiment of computational GI using a digital micromirror device. Here we present a rather different evaluation criterion named receiver operating characteristic (ROC) analysis that serves as the performance of merit for the quantitative comparison. A ROC curve is created by plotting the true positive rate against the false positive rate at various threshold settings. Both theoretical analysis and experimental results demonstrate that the ROC curve and the area under the curve are better and more intuitive indicators of the performance of the GI, compared with conventional evaluation methods. Additionally, for examining gray-scale objects, the calculation of the volume under the ROC surface is analyzed and serves as a performance metric. Our scheme should attract general interest and open exciting prospects for ROC analysis in thermal GI.

© 2015 Optical Society of America

OCIS codes: (120.4800) Optical standards and testing; (110.1650) Coherence imaging; (110.3000) Image quality assessment; (110.3010) Image reconstruction techniques; (000.5490) Probability theory, stochastic processes, and statistics.

http://dx.doi.org/10.1364/JOSAA.32.001084

1. INTRODUCTION

Ghost imaging (GI) is an amazing imaging technique that relies on the correlation between a pair of the outputs of two photodetectors: a bucket detector with no spatial resolution that collects the total intensity of light that has interacted with the object to be imaged, and a spatially resolving detector that is illuminated by light that has not interacted with the object. Neither detector’s output suffices to produce an image, but an image can be retrieved when we measure the cross correlation between these two outputs. The initial demonstration of ghost imaging [1,2] used entangled photons produced by spontaneous parametric downconversion to perform the imaging; hence it was thought that the origin of the phenomena was ascribed to the quantum entanglement of photons [3]. However, over the following years, a debate [4,5] on the big question of whether the physical origin of ghost imaging can be explained entirely using classical intensity correlations was soon sparked, when it was demonstrated both theoretically [6] and experimentally [7,8] that ghost imaging is also achievable with pseudothermal or true thermal light [9]. In 2012, another debate about the quantum versus semiclassical interpretations for pseudothermal ghost imaging was aroused and was finally settled [10–12].

Later in 2008, it was realized that the purpose of the reference beam was merely to measure the field distribution at the object, so it could just as well be replaced by a single predetermined spatially modulated object beam plus the bucket (single-pixel) detector. This modified technique was called computational GI by Shapiro [13]. Interestingly, in the same year, Baraniuk et al. proposed a single-pixel camera scheme [14] based on compressed sensing (CS) [15–17]. The only difference between the optical systems of a single-pixel camera and computational GI is that the spatial modulator was placed behind the object instead of in front. Later, Katz et al. [18] experimentally demonstrated that utilizing CS algorithms in GI could get a much better performance than conventional GI. Actually, since the measurement model is the same, both computational GI and single-pixel camera schemes can use either GI or CS reconstruction algorithms, and with the latter the number of measurements will be greatly reduced. Although they were historically independent and developed in parallel with each other, both techniques promise a resource-efficient alternative to array detectors like charge-coupled device (CCD) cameras, permitting us to reduce operational problems involved in systems based on raster scanning. Additionally, the intensity coming from image pixels is converged to a bucket (single-pixel) detector by a collecting lens,
thus obtaining a maximum flux and a great improvement of signal-to-noise ratio (SNR), as well as avoiding the average allocation of flux on the pixel dimension, compared with high-resolution array detection or raster scanning. In addition, they can form images at wavelengths for which high-quality array imagers are unavailable. It is worth mentioning that the array detection will instead outperform the single-pixel computational imaging scheme in terms of the speed of the image acquisition.

In 2012, Luo and co-workers [19,20] found a fascinating phenomenon that nonlocal imaging by conditional averaging of random reference measurements could improve the visibility and reduce the number of exposures and computation time. They called this technique correspondence imaging. In recent years, some other approaches have been developed to improve the quality of GI, such as differential ghost imaging [21,22], fast time-correspondence imaging [23,24], and adaptive compressive ghost imaging [25]. Based on the wide variety of variations on the GI theme, GI has been widely used in the fields of remote sensing [26], optical encryption [27,28], and lensless ghost imaging with sunlight [29].

Despite a large number of studies aiming at improving the performance of GI algorithms, such as SNR, the image contrast, and the image visibility, to our knowledge, there has been a lack of study on the unified evaluation criteria of these different algorithms used for GI. We believe that a study from such an angle will be helpful to elucidate whether a new algorithm is more effective than others in GI or not. In statistics, a receiver operating characteristic (ROC), or ROC curve [30–33], is a graphical plot that illustrates the performance of a binary classifier system as its discrimination threshold is varied. To our knowledge, ROC analysis has been already widely used in medicine [34], radiology [35], biometrics [36], data mining, and other areas. The ROC curve is created by plotting the true positive rate against the false positive rate at various threshold settings. It is known that GI often uses binary masks as objects; thus, ROC is very suitable for GI algorithm evaluation.

The purpose of the present paper is not to refute any of the image quality evaluation methods of GI, but instead to present a rather different estimation method applying the ROC curves. The paper is organized as follows. In Section 2, we first discuss the theory of the ROC, then describe how to use the ROC curve to evaluate the performances of the GI algorithms. After that, in Section 3, an experiment of computational GI is built and the results show that with this method we can make an intuitive comparison of the different reconstruction algorithms with the increase of the number of frames. And in Section 4, the conclusion is made.

2. THEORY OF THE RECEIVER OPERATING CHARACTERISTIC ANALYSIS

Now let us consider a two-class prediction problem (binary classification), in which the outcomes are labeled as either positive (P) or negative (N). For example, imagine that the test outcome positives and the test outcome negatives are normally distributed with different means respectively, as shown in Fig. 1. Certainly, they can also be in other value (or probability) distributions with some overlap. By using a binary classifier, four possible outcomes will appear. If the outcome from a prediction is P and the actual value is also P, then it is called a true positive (TP) or hit; however, if the actual value is N then it is said to be a false positive (FP). Conversely, a true negative (TN) or correct rejection has occurred when both the prediction outcome and the actual value are N, and false negative (FN) is when the prediction outcome is N while the actual value is P. For GI, here we only consider a binary object mask, either transparent (1) or opaque (0).

Let us define an experiment from P positive instances and N negative instances for some conditions. The four outcomes can be formulated in a $2 \times 2$ contingency table or confusion matrix, as given in Fig. 2, where Total population $= TP + TN + FP + FN$. The contingency table can derive several evaluation metrics defined as follows.

The true-positive rate (TPR) is also known as sensitivity (or called the probability of detection, to use radar terminology):

$$TPR = \frac{TP}{TP + FN}. \tag{1}$$

Specificity or true negative rate (TNR):

$$TNR = \frac{TN}{FP + TN}. \tag{2}$$

The false-positive rate (FPR) is also known as the fall-out and can be calculated as 1-specificity (or called the probability of false alarm, to use radar terminology):

\begin{table}[h]
\centering
\begin{tabular}{|c|c|c|}
\hline
\textbf{True status} & \textbf{Condition positive} & \textbf{Condition negative} \\
\hline
\textbf{Test outcome positive} & TP & FP (Type I error) \\
\textbf{Test outcome negative} & FN (Type II error) & TN \\
\hline
\textbf{Total} & TP+FN & FP+TN \\
\hline
\end{tabular}
\caption{$2 \times 2$ contingency table for each threshold.}
\end{table}
FPR = \frac{FP}{FP + TN} = 1 - \text{Specificity} = 1 - \text{TNR}. \hspace{1cm} (3)

Miss rate or false negative rate (FNR):
\[ \text{FNR} = \frac{FN}{TP + FN} = 1 - \text{TPR}. \hspace{1cm} (4) \]

To draw an ROC curve, the TPR (as the y-axis) and FPR (as the x-axis) are needed. Thus, the ROC curve is the sensitivity as a function of 1-specificity, which depicts relative trade-offs between true positive (benefits) and false positive (costs). Actually, the TPR (or FPR) defines how many correct (or incorrect) positive results occur among all positive (or negative) samples available during the test. Each prediction result or instance of a confusion matrix represents one point in the ROC graph. The point in the upper left corner or coordinate (0,1) of the ROC space represents 100% sensitivity (no false negatives) and 100% specificity (no false positives); that is, if a classifier algorithm reaches this point, then this algorithm can be thought to be the best possible prediction method. Thus, the (0,1) point is also a perfect classification. The diagonal line passing the (0,0) point with a slope of 1 is called the line of no-discrimination, which divides the ROC space. A completely random guess would give a point along this line. Points above (or below) the diagonal represent good (or poor) classification results, i.e., better (or worse) than the completely random guess.

3. EXPERIMENTAL ANALYSIS

A. Experiment of Computational Ghost Imaging Using a Digital Micromirror Device

The scheme of computational GI based on a digital micromirror device (DMD) is given in Fig. 3. In order to generate spatially random modulated patterns, we used a DMD instead of a spatial light modulator. The DMD is a 1024 × 768 array of micromirrors each of size 13.68 × 13.68 μm², and which can be oriented at +12° and −12° away from the initial position; thus, the light falling on it will be reflected into two directions. Micromirrors are all independently configurable at frequency up to 32 kHz.

In our experiment, as a light source, we used a halogen lamp of 55 W power. At the fiber output, the light beam first passed through an aperture diaphragm and a beam expander in order to be expanded into a 2.2 mm diameter collimated light and was then illuminated a 160 × 160 pixel region on the DMD. The micromirrors outside this region are kept in the −12° position. This beam was oriented toward the DMD at an angle of incidence corresponding to twice the tilting angle of the DMD micromirrors (approx. 24°). Micromirrors oriented at +12° would reflect the light and image the pattern onto the object through a 2f/2 lens imaging system (consisting of an imaging lens of aperture 25.4 mm and focal length 50.8 mm), and appear as bright pixels; inversely, micromirrors oriented at −12° appear as dark pixels. The patterns I_R encoded on the DMD are completely random and the spatial resolution of the pattern on the object plane is also 13.68 μm. Here we took a black-and-white film printed with “A” as the transmission object, either transparent (1) or opaque (0). Then the transmitted light was converged to a bucket (single-pixel) detector via a collecting lens, and the total light intensity was recorded as I_B. Here we used a 1/1.8 in. CCD with 1280 × 1024 pixels and an exposure rate of 26 frames/s as the bucket (single-pixel) detector, by integrating the gray values of all the pixels in each exposure.

A ghost image of the object can be reconstructed by cross correlating the random binary patterns I_R with the recorded total intensity S_B of the bucket detector:

\[ G^{(2)} = \langle S_B I_R(x_R) \rangle, \hspace{1cm} (5) \]

where \( \langle \cdot \cdot \cdot \rangle \) is an average operator and \( x_R \) denotes the spatial position in the reference frames.

The normalized second-order correlation function \( g^{(2)} \), which equals unity for coherent light [3], is given by

\[ g^{(2)} = \frac{\langle S_B I_R(x_R) \rangle}{\langle S_B \rangle \langle I_R(x_R) \rangle}, \hspace{1cm} (6) \]

and can be used to form a correlation-based ghost image as an alternative to the \( G^{(2)} \) approach.

The object information can also be extracted by the cross-covariance between the bucket and [6,8], which is given by

\[ \Delta G = \langle S_B I_R(x_R) \rangle - \langle S_B \rangle \langle I_R(x_R) \rangle. \hspace{1cm} (7) \]

If we use \( \langle S_B \rangle \) as a boundary, \( \{ S_B \} \) can be divided into two subsets: \( \{ B_+ | S_B > \langle S_B \rangle \} \) and \( \{ B_- | S_B < \langle S_B \rangle \} \). According to Luo’s method, after the partition, a positive (or negative) image can be retrieved by only averaging the frames corresponding to \( B_+ \) (or \( B_- \)), identified by the subscript “+” (or “-”): \( R_+ = \langle I_{R_+} \rangle, R_- = \langle I_{R_-} \rangle \). Then the correspondence image can be retrieved by

\[ C_I = R_+ - R_- = \langle I_{R_+} \rangle - \langle I_{R_-} \rangle. \hspace{1cm} (8) \]

The bucket signal can be also defined as \( S_B = \int_{A_{\text{beam}}} I_B(x_R) T(x_B) d^2x_B \), where \( x_B \) denotes the spatial position of the object, \( I_B \) is the intensity distribution on the object, \( T(x_B) \) represents the intensity transmission function of the object, and \( A_{\text{beam}} \) is the area of the beam. The reference bucket signal is defined as \( S_R = \int_{A_{\text{beam}}} I_B(x_R) d^2x_B \). In computational GI scheme, \( I_B(x_R) = I_B(x_R) \); thus, the total instantaneous transmission function of the object can be expressed as \( \tilde{T} = \langle S_B \rangle / \langle S_B \rangle \). The core idea of DGI is to use a differential bucket signal \( S_{\Delta} = \int_{A_{\text{beam}}} I_B(x_R) \Delta T(x_B) d^2x_B = S_B - \langle S_B \rangle S_R \), to replace original \( S_B \). Then a differential ghost image [22] can be obtained via
DGI = \langle \delta S \delta I_R(x_R) \rangle = \langle S_R I_R(x_R) \rangle - \langle S_R \rangle \langle I_R(x_R) \rangle. \tag{9}

Often, the object \( x \in \mathbb{R}^N \) (reshaped from a two-dimensional image matrix of size \( r \times c \) pixels, \( N = r \times c \)) can be well-approximated as a linear combination of just a few of the largest coefficients from a certain basis \( \Psi = [\psi_1, \psi_2, \ldots, \psi_N] \) (e.g., Haar wavelet basis, discrete cosine transform basis, or Fourier transform basis):

\[
x = \Psi x', \quad \text{or} \quad x = \sum_{i=1}^{N} x'_i \psi_i
\]

where \( x' \in \mathbb{R}^N \) is the coefficient sequence of \( x \). Mathematically, we say that a signal \( x' \) is \( k \)-sparse when it has at most \( k \) nonzeros, i.e., \( \|x'\|_0 = \lim_{\rho \to 0} \sum_{\|x'\|_\rho} = \|\{i : x'_i \neq 0, 1 \leq i \leq N\}\| \leq k \). Actually, this is a very simple and intuitive measure of sparsity of a vector \( x' \), directly counting the number of nonzero entries in it [37]. Additionally, GI systems typically acquire the bucket measurements in the form of inner products between the object signal and modulated patterns. Consider the general problem of reconstructing a vector \( x' \) from linear measurements \( y \) about \( x' \) of the form

\[
y = A \Psi x' + e, \tag{11}
\]

where \( y \in \mathbb{R}^M \) is a column vector consisting of the bucket signal, \( M \) is the total number of reference frames, \( e \in \mathbb{R}^M \) is the noise, and \( A \in \mathbb{R}^{M \times N} \) is a measurement matrix (each row of \( A \) records the brightness of pixels in one frame \( I_B \); i.e., each pattern on the DMD should be also flattened into a row vector, and then \( M \) binary patterns are rearranged row by row to form the measurement matrix \( A \)). At first glance, solving the underdetermined system of equations (\( M \ll N \), i.e., we have many fewer measurements than unknown signal values) appears hopeless, but we know that the signal \( x \) is sparse or compressible, meaning that it essentially depends on a number of degrees of freedom which is smaller than \( N \), and making the search for solutions feasible. In fact, accurate and sometimes exact recovery is possible by solving a simple convex optimization problem [17]. Recent research [38] has proved that the use of total variation (TV) regularization instead of the \( l_1 \) term in CS problems gives a sharper recovered image by preserving the edges or boundaries more accurately, and the gradient of an image is generally sparse as well. Here, for reconstructing an image, a solver named TVAL3 [38] is applied to this TV-based minimization model:

\[
\min_x \sum_i \|D_i x\|_1 + \frac{\mu}{2} \|y - Ax\|_2^2, \tag{12}
\]

where \( D_i x \) is the discrete gradient vector of \( x \) at position \( i \), \( D \) is the gradient operator, \( l_\rho \) norm is defined as \( \|x\|_\rho = (\sum_{j=1}^{\rho} |x_j|^\rho)^{1/\rho}, \|D_i x\|_1 \) is the discrete TV of \( x \), and \( \mu \) is a constant scalar used to balance these two terms. The first term is small when \( D_i x \) is sparse. The second term is small when the optimal \( x \) is consistent with Eq. (11) within a small error. Therefore, the fundamental idea behind CS is: rather than performing the process of massive data acquisition followed by compression, CS finds ways to directly sense the data in a compressed form, i.e., at a lower sampling rate.

B. Performance Evaluation of Different Reconstruction Algorithms

Here, we mainly made a comparison of the performances of these six GI reconstruction algorithms mentioned above. The imaging region is \( 160 \times 160 \) pixels; thus, \( N = 25,600 \). To demonstrate the effect of the number of measurements on the visibility, the experimental results of the same target with the number of frames changing from 597 to 7761 are given in the different columns of Fig. 4. Each row of Fig. 4 shows results retrieved by one algorithm. All the recovered images are normalized to a range of 0–1. Additionally, in order to allow a fair comparison of algorithms, the exposure time and the modulation frequency used for \( G^{(2)} \), \( g^{(2)} \), \( \Delta GI \), CI, DGI, and CS are the same. It can be seen that as the number of measurements increases, the image quality of these methods all becomes better. In addition, \( G^{(2)} \) is the worst algorithm, while CS (here we used TVAL3 algorithm) is relatively the best one for its enforcement of sparsity.

Now we perform a simple binary test for each image pixel to judge whether the value on this pixel of the reconstructed image is larger than the given threshold, i.e., \( H_0 \) (test outcome negative) is no “A” present and \( H_1 \) (test outcome positive) is “A” present at a particular location, scale, and orientation. As mentioned in Fig. 2, four possible test outcomes will be divided by each threshold, and can be formulated in a 2 \( \times \) 2 contingency table, where TP (or TN) is the number of pixels in the intersection of two support sets of test outcome positive (negative) and condition positive (negative), and then FP (FN) equals to the total number of test outcome positive (negative) minus

![Fig. 4. Comparison of experimental results reconstructed by different algorithms. Each row stands for images recovered by \( G^{(2)} \), \( g^{(2)} \), \( \Delta GI \), CI, DGI, and CS, respectively, while each column represents the number of patterns varying from 597 to 7761, with an equal-interval of 1791.](image-url)
TP (TN). The threshold can be changed in a value range \([\text{min(test outcomes), max(test outcomes)}]\). One can adjust the threshold (green vertical line in Fig. 2), which will in turn change the FPR and TPR. By traversing through a value range of thresholds, we will get a sequence of 2 × 2 contingency tables, from which a series of TPRs and FPRs will be computed. After all these have been done, the coordinate \((\text{FPR}, \text{TPR})\) corresponds to \(i\)th threshold, and all these coordinate points will form a ROC curve, where FPR serves as the \(-x\)-axis and TPR is treated as the \(y\)-axis. Furthermore, it can be easily found that increasing the threshold will result in fewer false positives (and more false negatives), corresponding to a leftward movement on the curve. Actually, the final shape of the curve is determined by how much overlap the two distributions have. Then we drew the ROC curves of these six methods, as illustrated in Figs. 5(a)–5(e), each graph corresponding to a different number of measurements. We can see that for the same number of measurements (or frames), the ROC curve of \(G^{(2)}\) is close to the line of no-discrimination; thus, the performance of \(G^{(2)}\) is the worst. The ROC curve of CI is above the diagonal line; thus, the performance of CI is much better than \(G^{(2)}\). The performance of \(\Delta G\) and DGI is similar, giving better classification results than that of CI. If we look carefully, we will notice that DGI gives a little better performance compared with the above four algorithms. And obviously, compressive GI is relatively the best choice for obtaining images with high visibility.

Generally, the peak signal-to-noise ratio (PSNR) and the mean square error (MSE) are used for a quantitative evaluation of the image quality:

\[
\text{PSNR} = 10 \log \frac{255^2}{\text{MSE}}, \tag{13}
\]

where \(\text{MSE} = \frac{1}{n} \sum_{i,j=1}^{n} (T(i,j) - \hat{T}(i,j))^2\), \(T\) represents the original image consisting of \(r \times c = N\) pixels, \(\hat{T}\) stands for the retrieved image, and \(255^2\) that appears on the right-hand side is due to the use of 8-bit amplitude quantization. Naturally, the larger the PSNR value is, the better the quality of the image recovered.

In GI, for binary masks, the contrast-to-noise ratio (CNR) \([39, 40]\) is often used as the figure of merit, which is defined to be

\[
\text{CNR} = \frac{\langle G(\vec{x}_{\text{in}}) \rangle - \langle G(\vec{x}_{\text{out}}) \rangle}{\sqrt{\frac{1}{2} \Delta^2 G(\vec{x}_{\text{in}})^2 + \Delta^2 G(\vec{x}_{\text{out}})^2}}, \tag{14}
\]

where \(\Delta^2 G(\vec{x}) = \langle G(\vec{x})^2 \rangle - \langle G(\vec{x}) \rangle^2\). Here \(\vec{x}_{\text{in}}\) and \(\vec{x}_{\text{out}}\) represent the pixel positions inside and outside the transmitting regions of the object. The variances \(\Delta^2 G(\vec{x}_{\text{in}})\) and \(\Delta^2 G(\vec{x}_{\text{out}})\) are generally not the same.

From the definition, we can see that PSNR, MSE, and CNR all quantify the visibility via the calculation of pixel errors. Since each natural image has a certain structure, the preceding pixelwise performance measures fail to capture the correlation structure of the image and may cause evaluation misjudgments; e.g., an image which is supposed to have a better visibility may instead have a worse PSNR, MSE, or CNR value.

In order to make up for the flaw, we introduce the area under the curve (AUC) \([34]\) as the evaluation criterion, which is defined as the integration of the TPR with spacing increment FPR. When using normalized units, a reliable AUC can be interpreted as the probability that the classifier will assign a higher score to a randomly chosen positive example than to a randomly chosen negative example. Since the AUC is a portion of the area of the unit square, its value will always be between 0 and 1. However, because random guessing produces the diagonal line between \((0,0)\) and \((1,1)\), which has an area of 0.5, no realistic classifier should have an AUC less than 0.5. Therefore we can use 0.5 as an evaluating baseline here. Here, the AUC results for each ROC curve in Fig. 5 are calculated and listed in Table 1. These AUC results are in agreement with the trend determined by the ROC curves. Therefore, there is no doubt that AUC can be treated as an evaluation criterion of the performances of the classifier algorithms (e.g., the GI reconstruction approaches for binary objects).

We found that AUC can not only evaluate the quality of the algorithms, but also estimate the quality of reconstructed images. To demonstrate the latter feature, we plotted the ROC curves of compressive GI with the increase of the number of measurements, as shown in Fig. 6, and listed the AUC estimate in Table 2. As the number of measurements (frames) increases, the ROC curve becomes steeper, close to the coordinate \((0, 0)\) of the ROC space, i.e., close to 100% sensitivity and 100% specificity; and the AUC also gets larger, only exhibiting a little slight oscillation. This trend seems in conformity with the theory. So AUC analysis can replace the role of traditional PSNR, MSE, or CNR.

Furthermore, Fig. 7(a) shows a basic ROC curve of DGI with the number of accumulated measurements of 7761. We find a cutoff point for best sensitivity and specificity (red circle in plot), corresponding to the threshold 0.48. By contrast, a cutoff point for best PSNR (green circle in plot) is also calculated, corresponding to the threshold 0.66. A photo of the object is taken by a CCD as the reference, presented in

![Fig. 5. ROC curves of different reconstruction algorithms with a number of measurements of (a) 597, (b) 2388, (c) 4179, (d) 5970, (e) 7761.](image)
The original result recovered by DGI is given in Fig. 7(c). The segmentation results divided by these two thresholds are separately shown in Figs. 7(d) and 7(e). However, in the pursuit of the best PSNR or MSE, it is essential to realize that the sensitivity and specificity under this threshold are also the best, so blindly running after the PSNR or the MSE will be one-sided. Actually, the cutoff point for best sensitivity and specificity is found by calculating the least distance \( d = \sqrt{\text{FPR}^2 + (1 - \text{TPR})^2} \), i.e., computing the closest point to the coordinate (0, 1) of the ROC space. Thus, the ROC curves of two different algorithms will be separate from each other to maximum extent under this closest points. Therefore, using the cutoff point for best sensitivity and specificity to evaluate the stand or fall of algorithms is more appropriate than traditional evaluation methods, which is another core advantage of the ROC analysis.

To our knowledge, ROC analysis is generally used for dichotomous diagnostic tasks. Since binary masks are widely used as objects in thermal GI, here we transplant ROC analysis to the performance evaluation of GI reconstructions. Although binary masks are easy to construct and they lead to the best ghost imaging results, such masks do not represent objects of interest in practical applications of GI. Therefore, we want to extend ROC curves for the examination of a broader class of objects, e.g., gray-scale objects. For such objects, ROC curves should deal with classification problems with more than two classes, which has always been cumbersome, as the degrees of freedom increase quadratically with the number of classes, and the ROC space has \( b(b - 1) \) dimensions, where \( b \) is the number of classes.

For simplicity and without loss of generality, we draw an image with three gray values (230, 98, and 0, with a pixel size of 4681, 8271, and 12,648, denoted as Class 1, Class 2, and Class 3, respectively) as the gray-scale complex digital object, as shown in Fig. 8(a), and use 6400 random patterns of the DMD for DGI and CS reconstructions, as shown in Figs. 8(b) and 8(c). The ROC space for such particular case with three classes (also called three-way ROC [42]) has six dimensions, and then with any three-dimensional coordinates we can plot a ROC surface. In such a trichotomous (three-alternative) diagnostic task, one can depict diagnostic performance with a \( 3 \times 3 \) contingency table, as illustrated in Table 3, where \( \pi = v_1 + v_2 + v_3 = r_1 + r_2 + r_3 \). Following a threshold judgment role, one can potentially misidentify values of 230, 98, and 0 in two ways, yielding six possible kinds of diagnostic errors. \( x_{11} \) is the number of pixels in the intersection of two support sets of Class 1’s test result and true status; \( y_{22} \) and \( z_{33} \) can be obtained in the same way. For each two thresholds \( \theta_1 \) and \( \theta_2 \) (0 < \( \theta_1 < \theta_2 < 255 \)), we set \( a = x_{12} + x_{13} = r_1 - x_1 \), \( b = y_{21} + y_{23} = r_2 - y_2 \), \( c = z_{31} + z_{32} = r_3 - z_3 \), and then we will compute

\[
\begin{align*}
\text{TPR}_1 &= \frac{x_{11}}{x_1}, \\
\text{FPR}_1 &= \frac{x_{12}}{x_1}, \\
\text{TPR}_2 &= \frac{a}{r_2}, \\
\text{FPR}_2 &= \frac{b}{r_2}, \\
\text{FPR}_3 &= \frac{c}{r_3}, \\
\text{TPR}_3 &= \frac{a}{r_2}. 
\end{align*}
\]

After having traversed the whole sets of double thresholds, we can draw a three-dimensional ROC surface based on the data, as shown in Figs. 8(d) and 8(e), where \( \text{TPR}_2 \), \( \text{TPR}_3 \), and \( \text{TPR}_1 \) are treated as the \( x, y, z \)-axes, respectively, based on the theory of three-way ROC [42]. The calculation of the volume under the ROC surface (VUS) \( \int_0^1 \int_0^1 \int_0^1 \text{f}(x, y, z) \text{d}x \text{d}y \text{d}z \) is analyzed and serves as a performance metric. Actually, VUS equals the probability that test values will allow a decision maker to correctly sort a trio of items containing a randomly selected member from each of three populations. From the data, we can see the VUS of CS is much higher than that of DGI, the same trend with the image quality.

### Table 1. AUC Estimate of Results of Different Algorithms for Different Numbers of Frames

<table>
<thead>
<tr>
<th>M</th>
<th>1000</th>
<th>2000</th>
<th>3000</th>
<th>4000</th>
<th>5000</th>
<th>6000</th>
<th>7000</th>
</tr>
</thead>
<tbody>
<tr>
<td>CS</td>
<td>0.9483</td>
<td>0.9647</td>
<td>0.9763</td>
<td>0.9784</td>
<td>0.9819</td>
<td>0.9911</td>
<td>0.9891</td>
</tr>
</tbody>
</table>

Fig. 6. ROC curves of CS algorithm (TVAL3) with the number of measurements increasing from 1000 to 6000.

Fig. 7. ROC analysis. (a) is the ROC curve of DGI using 7761 patterns. (b) is the original image. (c) is the image reconstructed by DGI. (d) and (e) are threshold segmentation results for the thresholds 0.48 and 0.66, corresponding to the red circle and the green circle in (a), with a PSNR of 7.9331 and 15.5637 dB, respectively.
Table 3. 3 × 3 Contingency Table for Three-Way ROCs

<table>
<thead>
<tr>
<th>Test Result</th>
<th>Class 1</th>
<th>Class 2</th>
<th>Class 3</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class 1</td>
<td>$x_{11}$</td>
<td>$x_{12}$</td>
<td>$x_{13}$</td>
<td>$r_1$</td>
</tr>
<tr>
<td>Class 2</td>
<td>$y_{21}$</td>
<td>$y_{22}$</td>
<td>$y_{23}$</td>
<td>$r_2$</td>
</tr>
<tr>
<td>Class 3</td>
<td>$z_{31}$</td>
<td>$z_{32}$</td>
<td>$z_{33}$</td>
<td>$r_3$</td>
</tr>
<tr>
<td>Total</td>
<td>$v_1$</td>
<td>$v_2$</td>
<td>$v_3$</td>
<td>$\pi$</td>
</tr>
</tbody>
</table>

The evaluation of reconstruction performance for the objects with more tones (i.e., multiclass problems) can also benefit from this method. As we know, there are some other approaches [32,34] based on an extension of AUC; thus, ROC analysis is powerful for performance evaluation of GI reconstruction algorithms.

4. CONCLUSION

In summary, since binary masks are widely used as objects in thermal GI, a ROC curve (created by plotting the true positive rate against the false positive rate at various threshold settings) is introduced here as a rather different evaluation criterion of performances of some GI algorithms. We have performed an experiment of computational GI based on a DMD. The performances of thermal ghost imaging, including traditional second-order correlation function, normalized GI, background-subtracted GI, GI, DGI, compressive GI, were compared in this paper. Both theoretical analysis and experimental results demonstrated that the ROC curve and the AUC are better and more intuitive indicators of the performance of thermal GI, compared with conventional evaluation methods. Additionally, for examining gray-scale objects, the calculation of the volume under the ROC surface is analyzed and serves as a performance metric.

Thus, our work is instructive and meaningful to the ROC analysis in thermal GI.

National High Technology Research and Development Program of China (2013AA122902); National Key Scientific Instrument and Equipment Development Project of China (2013YQ030595).

We are grateful to Wei-Mou Zheng for his illuminating discussions and comments on the topic presented here. We warmly acknowledge Ling-An Wu for providing suggestions, encouragement, and helpful discussions.

REFERENCES


