Assessment of speckle denoising filters for digital holography using subjective and objective evaluation models

ELSA FONSECA[2,3,*], PAULO T. FIADEIRO[2,3], MARCO V. BERNARDO[1], ANTÓNIO PINHEIRO[1,3], AND MANUELA PEREIRA[1,3]

[1] Instituto de Telecomunicações (IT)
[2] Fiber Materials and Environmental Technologies (FibEnTech)
[3] Universidade da Beira Interior (UBI)
[*] Corresponding author: efonseca@ubi.pt

Compiled September 12, 2019

Digital holography is an emerging imaging technique for displaying and sensing three dimensional objects. The perceived image quality of a hologram is frequently corrupted by speckle noise due to coherent illumination. Although several speckle noise reduction methods have been developed so far, there are scarce quality assessment studies to address their performance and they typically focus solely on objective metrics. However, these metrics do not reflect the visual quality perceived by a human observer.
In this work, the performance of four speckle reduction algorithms, namely the nonlocal means, the Lee, the Frost and the block matching 3D filters, with varying parameterizations, were subjectively evaluated. The results were ranked with respect to the perceived image quality to obtain the mean opinion scores using pairwise comparison. The correlation between the subjective results and twenty different no-reference objective quality metrics was evaluated.
The experiment indicates that block matching 3D and Lee are the preferred filters, depending on hologram characteristics. The best performing objective metrics were identified for each filter.

http://dx.doi.org/10.1364/ao.XX.XXXXXX

1. INTRODUCTION

Digital holography is a three-dimensional imaging technique that offers unique parallax and field of view capabilities, as well as 3D shape recovery of both macroscopic and microscopic objects. A wide range of applications of scientific nature includes industrial nondestructive evaluation, solid modeling, biomedical research, machine vision, among others. Recent advances in digital image sensors, computational equipment, holographic displays and 3D rendering techniques also offer new opportunities in the entertainment field.

In a digital hologram, the wave field transmitted or scattered by an object is encoded in the form of an interference pattern with a reference beam produced by a coherent light source, typically a laser. Due to the coherent nature of the laser light, a signal dependent multiplicative noise called speckle occurs. Speckle noise degrades the image quality of numerically reconstructed digital holograms in conventional 2D displays as well as the optical quality in holographic displays, imposing severe limitations in spatial resolution, signal-to-noise ratio, and phase accuracy. Among all noise contributions that deteriorate the quality of a digital hologram, speckle noise seems to have the most hindering effect and is difficult to remove by traditional filtering techniques.

The development of filtering techniques for speckle denoising of digital holograms is a very active research area. However, only a few comparative studies and comprehensive reviews have been published so far. A very thorough review was recently provided by Bianco *et al.*[1], where the available methods are classified into 2 main groups: optical or numerical. The first is performed during the acquisition process. Examples are averaging by wavelength [2] and phase diversity [3], combining different polarization states or slightly shifting or rotating the object. The second are applied to the reconstructed holograms using signal processing techniques and can be further divided

into spatial domain and transform domain filters. Optical based methods are out of the scope of this study since any comparison would require a large amount of hardware resources.

Few studies comparing the performance of speckle filters have been published so far. Wavelet-based techniques were assessed by K. M. Molony *et al* [4] in a study that addressed the behavior of stationary wavelet transforms and discrete wavelet transforms, with soft and hard thresholding, in terms of speckle suppression effectiveness and edge preservation capability. After testing 51 wavelets on phase-shifting and off-axis digital holograms, they found the best results for the Haar mother wavelet with soft thresholding at decomposition level 5, outperforming more traditional approaches such as mean filtering, median filtering, and Fourier filtering. Srivastava *et al* [5] compared 7 filters against their proposed partial differential equation based homomorphic method, using different objective reference metrics. Recently, 34 speckle reduction filters were tested on digital holographic phase imaging by S. Montresor and P. Picart [6]. The individual filters were tested on simulated phase fringe patterns according to 4 quality reference metrics: the gain in signal-to-noise ratio, the Q index, the standard deviation of the phase error and the signal-to-distortion ratio.

The perceived degradation effect of speckle noise in numerical reconstructions of digital holograms is not easily related with the quality metrics that are commonly used in 2D images. In a study by Lehtimäki *et al* [7], it is even suggested that, when holograms are viewed in autostereoscopic displays, the perceived speckle level is reduced and the apparent depth of field is increased. The perceptual quality of digital holograms viewed on stereoscopic displays has been inspected with respect to noise, blur, and perceived depth [8]. In another study, the same authors addressed the problem of comparing the error introduced by lossy compression on reconstructed holograms with the corresponding subjective quality findings [9]. They used mean filtering and Fourier filtering to reduce the negative impact of speckle noise in compression results. Darakis *et al* [10] addressed the effects of compression on reconstruction quality of highly speckled experimental holograms. Ahar *et al* [11] proposed a subjective quality assessment of 2D reconstructions of computer-generated holograms (CGH) has been performed for the comparison of compression techniques. They used the reference metrics peak signal-to-noise ratio (PSNR) and the structural similarity index measure (SSIM) and concluded that these have good correlation with the perceived visual quality. More recently, the authors proposed a versatile similarity measure (VSM) for complex amplitude holographic data. They compared its predictor performance, along with that of two traditional reference metrics, the mean square error (MSE) and the SSIM, against the provided mean opinion scores (MOS) of reconstructed computer generated holograms [12]. Later on, the VSM was tested in combination with a new full reference metric named sparseness significance ranking measure (SSRM) [13]. The later is based on sparse coding and a ranking system for the magnitudes of the spatial frequency coefficients and, either used alone or in combination with VSM, has proved to be more effective than some traditional 2D metrics when applied to CGHs. Recently, the group of T. J. Naughton addressed the loss in quality perceived by human observers of numerical reconstructions of compressed phase-shifting digital holograms has been related to MSE [14]. The authors found that the referred metric does not resemble the perceptual quality of the compressed hologram, possibly due to changes in speckle noise characteristics resulting from lossy compression.

This paper is organized as follows: section 2 describes the basics of digital hologram reconstructions and speckle noise characteristics. In this section, the filtering methods used for speckle denoising are also briefly reviewed. Section 3 presents the objective quality metrics for the evaluation of the selected filters. The subjective assessment methodology, including the experimental design implementation and the statistical framework, is discussed in section 4. Section 5 reports and analyses the subjective assessment results, while section 6 draws the concluding remarks.

2. SPECKLE DENOISING

In this work, a subjective assessment approach is put forward, aiming at understanding the perceptual characteristics of enhanced holograms. This procedure is followed by an extensive evaluation of the image quality of denoised reconstructions using objective image quality metrics (IQM). The proposed methodology was tested on the amplitude component of numerical reconstructions of digital holograms from the EmergImg-HoloGrail database[1], available online. This database comprises several experimental holograms acquired with a 4-step phase-shifting optical setup. Each experimental hologram results from an algebraic combination of 4 phase-shifted interferograms that is subsequently reconstructed to yield the amplitude of the object field. The Fresnel Transform Method is used for the numerical reconstructions. The description of the optical setup and of the numerical reconstruction method can be found in references [15, 16].

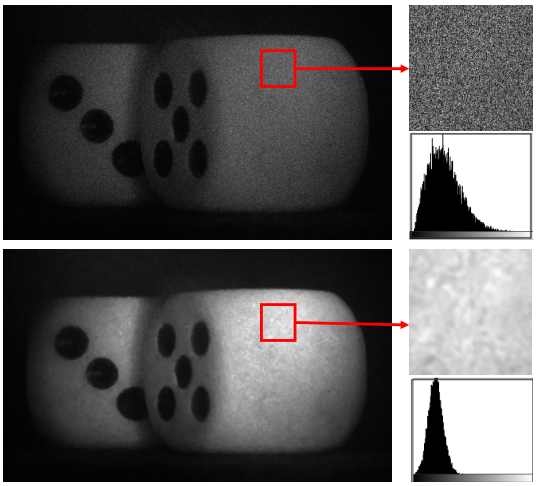


Fig. 1. Reconstruction of the Dice2 hologram and the histogram of a cropped area before (upper image) and after (lower image) filter processing.

The speckle patterns are a consequence of the large coherence length of the recording laser light source and the roughness of the object surface. Due to this roughness, each point at the object surface introduces a slight random phase delay in the corresponding scattered wavelet which, upon interference with the reference wave, results in an irregular pattern of maxima and minima. The complex speckle field scattered by the object's surface has a Gaussian probability density function with zero mean for its real and imaginary parts. As a consequence, the amplitude is Rayleigh distributed and the phase has uniformly distributed [17, 18]. When considering the intensity in

[1] http://emergimg.di.ubi.pt

the hologram plane, and in the case of fully developed speckle, a negative exponential is obtained. The same holds for the reconstruction plane and the corresponding probability density function is given by

$$p_I(I) = \frac{1}{2\sigma^2} \exp\left(-\frac{I}{2\sigma^2}\right), \quad \textbf{(1)}$$

where, $I \geq 0$, and $\sigma^2 = \langle I \rangle / 2$ with the brackets meaning the ensemble average of its argument.

Speckle noise is multiplicative and signal dependent, thus requiring sophisticated filtering approaches as compared to simple additive noise. However, it is common to perform an averaging procedure during the acquisition process, thus changing the speckle statistics to a Gamma distribution that tends to a Gaussian distribution, as the number N of samples increases

$$p_I(I) = \left(\frac{N}{\langle I \rangle}\right)^N \frac{1}{\Gamma(N)} I^{N-1} \exp\left(-\frac{N}{\langle I \rangle} I\right). \quad \textbf{(2)}$$

The holograms used in this work result from an averaging procedure with 15 samples. Figure 1 illustrates the effect of the BM3D filter on the speckle appearance and histogram of a homogeneous region selected from a reconstruction of a digital hologram.

As previously mentioned, this work addresses digital strategies to tackle the denoising problem, since optical approaches are much more complex to compare as they require different optical setups. Because subjective evaluation methods can be tiresome, the number of filters used in this comparative study must be kept at a modest level in order to keep the duration of the test within reasonable limits. Bearing this in mind, we selected 4 out of the most effective filters among those that have been more frequently tested on amplitude digital holograms. Specifically, 2 state of the art filters were selected: the block matching 3D approach [19, 20] and the non-local means filter [21]. Additionally, 2 traditional techniques for speckle removal in a variety of imaging modalities (such as Synthetic Aperture Radar and Ultrasound) that usually behave well in DH were chosen: the Lee [22] and the Frost [23] filters. This selection took into account a preliminary inspection of several speckle filters selected from an extensive study by S. Montresor and P. Picart [24] using objective metrics. According to the rank produced by this study, the top 10 positions were occupied by the block matching 3D, non-local means filter, Frost and Lee filters, as well as a series of wavelet based filters that didn't work specially well on our database.

Next, we give a brief description of the selected filters.

1) BM3D. The block matching 3D filter (BM3D) is an advanced filter that uses transform-domain shrinkage operators and collaborative filtering in combination with the decomposition of the image in patches, similarly to the non-local means filter. The BM3D filter promotes sparsity in the transform-domain by grouping similar 2D image blocks into 3D data arrays with high level of correlation, which are called groups. This operation is followed by 3 successive steps: 3D transformation of a group, shrinkage of the transform spectrum, and inverse 3D transformation. The shrinkage operation on the transform coefficients effectively attenuates the noise. The subsequent inverse 3D transform yields an estimate that consists of the jointly filtered grouped image blocks. After returning to their original positions, these blocks provide several different overlapping estimates for each pixel that need to be combined appropriately to obtain the final image. This is achieved by a procedure called aggregation, through which a weighted average of all overlapping blocks is obtained.

In this work, a BM3D algorithm available online [19] was used. Although this algorithm was designed for additive white Gaussian noise, it has been successfully tested on digital holograms by Montresor et al [25]. Similarly to the latter authors, we have decided to keep the default parameters given with the code, except for the selected optimization parameter that in this case corresponds to the standard deviation σ of the Gaussian noise.

2) NLM. The non-local means (NLM) method is a recently proposed filter, known for its remarkable ability to preserve image details while efficiently removing white noise [21]. It has been recently adapted to speckle noise in ultrasound images [26] and also tested for despeckling digital holograms [27]. The NLM explores a new patch-based non-local recovery paradigm, according to which the comparison of a given pixel with its neighboring pixels is replaced by comparison with non-local patches. The idea is that an estimate of a pixel located in a square patch may be obtained from the correlations with patches of the same size, having similar structures, located anywhere in a selected search window. The latter can be as large as the whole image, however, as its size increases, the computation cost increases dramatically. There are 3 parameters to consider in the NLM filter, namely, the size of the search window, the size of the similarity window, and finally h, which sets the strength of the filter, has to be adapted to the noise level of the particular image. In this work, the 3 parameters were used for optimization.

3) Lee. Synthetic Aperture Radar images also suffer from speckle noise, so, in principle, filtering techniques devised for this type of images can be easily adapted to digital holograms. Lee proposed an adaptive filter [22] that modifies its parameters according to the local statistical properties of the image to be restored. The estimation of these parameters is performed in a window centered on the pixel (i, j) to be restored according to the formula

$$d(i,j) = \mu_s(i,j) + \frac{\sigma_s^2(i,j)}{\sigma_s^2(i,j) + \sigma_n^2(i,j)} \left[s(i,j) - \mu_s(i,j)\right], \quad \textbf{(3)}$$

where, $d(i,j)$ is the denoised image, $s(i,j)$ is the noisy image, $\mu_s(i,j)$ is the local mean of the signal, $\sigma_s^2(i,j)$ is the local signal variance, and $\sigma_n^2(i,j)$ is the local noise variance. A local average operator can be implemented by a normalized low-pass filter, where the support region of the filter is defined by the size of the window selected for the estimation of local parameters. In this paper, the window size was used as an optimization parameter for the Lee filter.

4) Frost. The Frost filter [23] is another locally adaptive algorithm that consists of an adaptive Wiener filter. It convolves the noisy image pixel values $s(i,j)$ within a fixed size window with an exponential decay kernel of size n. The estimated pixel value is computed by

$$d(i,j) = s(i,j) \otimes m(i,j), \quad \textbf{(4)}$$

where, $s(i,j)$ is the noisy image, and the damping kernel is given by $m(i,j) = \kappa\alpha \exp\left(-\alpha \left|t\left(i,j;i_0,j_0\right)\right|\right)$. Here, κ is a normalization factor, $\left|t\left(i,j;i_0,j_0\right)\right|$ is the distance between the current pixel and a pixel located at point (i_0, j_0). Moreover, $\alpha = \left(4/\left(n\overline{\sigma}\right)\right)\left(\sigma_s^2/\mu_s^2\right)$ is a damping factor where σ_s is the local variance, μ_s is the local mean, and $\overline{\sigma}$ is the image coefficient of variation value. Herein, the size n of the kernel window is used as an optimization parameter.

3. PERCEPTUAL EVALUATION

The performance of different speckle reduction algorithms is subjectively assessed in 2-phases experiment. Since these are parametric algorithms, the parameters optimization was considered in phase 1. The optimal parameter found in phase 1 was used and the different speckle reduction algorithms were compared in phase 2.

A. Stimuli Generation

Four holographic images were selected for the test session and one for the training session. All the holographic images were selected from the EmergImg-HoloGrail database[2]. The corresponding 2D numerical reconstructed amplitudes are represented in Figure 2. The first 4 images correspond to the holograms used for the testing session (Astronaut, Car, Dice1, Skull) and the last one correspond to the hologram used for the training session (Dice2).

Speckle reduction using 4 different filters, and a set of 5 different filter parametrization levels was applied to each original 2D numerical reconstructed amplitude. The applied filters were BM3D, Frost, Lee, and NLM, described in section 2. The 5 different filter parametrization levels are indexed from 0 to 4. The index 0 corresponds to the raw speckled image and the remaining indexes correspond to the parameters presented in Table 1.

Table 1. Filter parameters.

Parameter index	BM3D (σ)	NLM (h, α, M)	Lee (n)	Frost (n)
1	1	0.25, 3, 5	3	3
2	2	0.50, 3, 5	5	7
3	5	1.00, 3, 5	9	11
4	20	1.00, 5, 9	15	17

From the 4 different holograms, the 4 different filters and the 5 different filter parametrization levels were generated a total of 80 stimuli. The generated stimuli represent different levels of speckle reduction, which is controlled by appropriate parametrization of the filters. These are the stimuli of phase 1 experiment.

Due to the duration of the experiments, phase 1 was divided into 2 parts, A and B. Part A concerns 4 holographic images and 2 filters (BM3D and Frost), and part B concerns the same 4 holographic images and the other 2 filters (Lee and NLM).

For phase 2, each 2D numerical reconstructed amplitude was filtered for speckle reduction using the same 4 filters used in phase 1 and considering the corresponding best filter parameter found in phase 1. A total of 16 stimuli were generated and considered for phase 2 experiment.

All stimuli of the subjective test can be found online[3].

B. Subjective Assessment

B.1. Methodology

The pair comparison (PC) method were used for both phases of this subjective test. It was considered the most appropriate for this study since it is a better representation of the process followed by humans when they have to decide between two products [28]. In this PC method a random pair of stimuli were simultaneously displayed side by side on the left (L) and right (R) sides of the screen monitor, according to the scheme presented in Figure 3, and a temporal two-alternative forced choice paradigm was used to collect the participant's selections.

The 5 different stimuli generated from a holographic image and a given filter are all compared to each other and considering different the comparisons left-right and right-left. Hence, a total of 320 comparisons were assessed, 160 comparisons in each part of phase 1. For phase 2, the 4 stimuli generated from a holographic image and the 4 filters, using the best parameter found on phase 1, are also compared to each other. Hence, a total of 48 comparisons are performed in this phase.

B.2. Test environment

The training and testing sessions were conducted in a test laboratory, which fulfills the recommendation for subjective evaluation of visual data issued by ITU-R [29]. The test laboratory room is equipped with a dim lighting system of 5500K of correlated color temperature. The color of all background walls and curtains in the room is mid gray. The laboratory setup is intended to ensure the reproducibility of the subjective test results by avoiding an unintended influence of external factors.

A MATLAB® application using a Psychophysics Toolbox was used to display the test stimuli on a full UHD 28″ ASUS PB287Q 4k monitor . The monitor has the following specifications: full resolution 4k (3840 $\times$ 2160 pixels), maximum luminance 300 cdm^{-2}. In the experiments, the background luminance around the stimuli corresponds to the mean luminance of the stimuli which is 30 cdm^{-2}. The ambient illumination did not directly reflect off of the monitor. The participants were seated in front and perpendicular to the center of the monitor at a distance of about 6$\times$ the stimuli height (1346 pixels, 21.5 cm), as suggested in recommendation ITU-R BT.2022 [30].

B.3. Participants

On phase 1 a total of 32 individuals participated in the test, 16 in part A and the other 16 in part B. The individuals that performed the assessment for part A were 9 male and 7 female with mean age of 23.9 $\pm$ 3.8 years old, while the assessment for part B was performed by 8 male and 8 female with mean age of 24.3 $\pm$ 4.3 years old. For phase 2, a total of 16 different individuals were considered, there were 9 male and 7 female with mean age of 25.1 $\pm$ 5.3 years old.

Before the experiments, an informed consent form was handed to the participants for signature and oral instructions were provided to explain the evaluation task. All subjects were screened for correct visual acuity using a Snellen acuity chart.

B.4. Test

The stimuli observation takes 5 seconds. After that, the participant has to choose the preferred stimulus (L or R) by pressing the 'left' or 'right' arrows on the keyboard to register the preference. The participant chooses L if left stimulus is the preferred one ($L > R$), or chooses R if right stimulus is the preferred one ($R > L$). The condition L similar to R, ($L = R$), is not considered in this study. It continues until the end of the session is reached. The participant's preferred scores were registered in participant matrices.

To reduce contextual effects, the order in which stimuli were displayed on the monitor screen was randomized differently for each participant and the same stimuli content was never shown consecutively.

[2] http://emergimg.di.ubi.pt
[3] http://emergimg.di.ubi.pt/Percep_Eval.html

(a) *Astronaut* (b) *Car* (c) *Dice1* (d) *Skull* (e) *Dice2*

Fig. 2. Experimental acquired holograms from EmergImg-HoloGrail-v2.

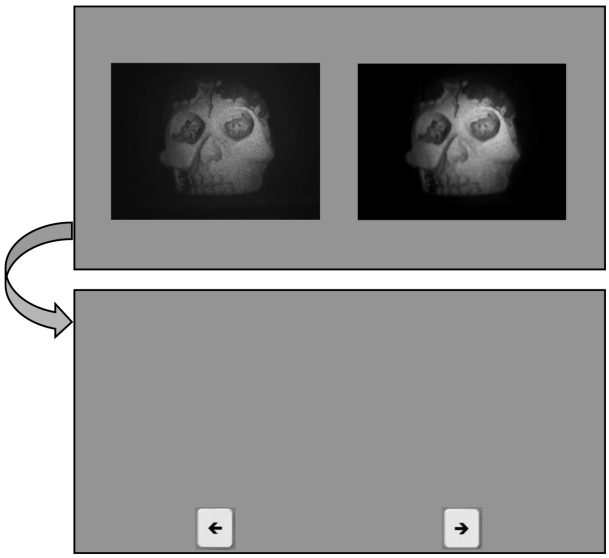


Fig. 3. Stimuli visualization sequence; Up: observation; Down: participant decision.

A training session was organized to allow the participants to get familiar with the evaluation procedure.

In each part of phase 1 experiment, the participant takes around 22 minutes to conclude the task including the training and testing sessions. Phase 2 was much shorter, taking around 7 minutes.

With all the participant's data, the MOS was computed for the parametrization levels of the filters and holographic images (phase 1), and also for the filter's performance on each holographic image (phase 2).

C. Data Analysis

The analysis and comparisons of subjective data is, in general, performed using the MOS. The two-alternative forced choice paradigm used does not provide directly such scores, and therefore the MOS values has to be estimated. The models of Thurstone and Bradley and Terry [28, 31] are two of the most used models. Since the differences between them are not relevant [32], in this case a variant of the Thurstone Case V model [28, 32] was used. A more detailed description of this model is provided in [32].

Before MOS estimation, and according to recommendation ITU-R.500 [29] for outlier detection, the subjective preferred results of the two-alternative forced choice experiments were firstly processed by detecting and removing participants whose preference scale scores deviate from others.

The MOS estimation consists of a continuous scale quality score computed from the binary scale ($L > R$ or $L < R$) that were previously registered in the participant matrices.

For phase 1 will be generated a total of 16 participant matrices, that corresponds to the 4 holographic images times 4 filters. Each matrix is 5×5 corresponding to the 5 parameter indices. For phase 2 will be generated a total of 4 participant matrices, that corresponds to the 4 holographic images. Each matrix is in this case 4×4 corresponding to the 4 filters.

Matrices with the winning frequencies w_{ij} have to be computed from the obtained participant matrices for each pair of stimuli i and j. Note that $w_{ij} + w_{ji} = N$, where N is the number of participants. Using the winning frequencies w_{ij} a count matrix C_{ij} has to be constructed. Each element of the count matrix C_{ij} is computed using equation (5)

$$C_{ij} = w_{ij} + 1, \tag{5}$$

where a prior of 1 was added on all the counts to regularize the estimates, meaning that a priori all choices are possible, which corresponds to Laplace smoothing [32]. Hence, C_{ij} represents the number of times stimulus i is preferred over stimulus j, where i and j are the row and column of the matrix, respectively.

The Thurstone Case V model [28] is used to estimate the quality score value μ_i for stimulus i. If only two stimuli, L and R, are compared, the quality score values (stimulus difference) are given by equation (6)

$$\mu_L - \mu_R = \phi^{-1}\left[\frac{C_{LR}}{C_{LR} + C_{RL}}\right], \tag{6}$$

where, ϕ^{-1} is the inverse cumulative distribution function of the standard normal. Then, applying the inverse cumulative distribution function to the data on this matrix one can obtain the quality score values. Thus, the estimated MOS simply corresponds to the matrix rows summation.

All results of the subjective test can be found online[4].

4. OBJECTIVE EVALUATION

The objective quality measures considered in this work apply to the absolute value of a 2D numerical reconstruction of a digital hologram at the considered focus distance. They are no-reference

[4] http://emergimg.di.ubi.pt/Percep_Eval.html

(NR) metrics and are divided in 2 main categories: speckle suppression and general quality measures. Metrics pertaining to the first group are applied to homogeneous areas, while the second apply to the whole image.

A. Speckle Suppression Metrics

Some metrics are specific to the nature of speckle noise and are traditionally used in the context of synthetic aperture radar imaging (SAR), although they have been tested in digital hologram reconstructions as well. These metrics are computed as the mean of the values obtained for a given statistics evaluated at selected homogeneous regions of the filtered image. Since the selection is difficult to perform automatically, several small areas are picked up manually. The procedure is applied to each separate reconstructed image, ensuring similar patch sizes, with enough number of pixels for statistical evaluation, and similar gray levels.

The speckle contrast (SC) [17] quantifies the level of speckle noise present in an image. It is defined by the ratio of standard deviation to the mean gray level in an homogeneous area of a given filtered image I_f

$$\mathrm{SC} = \frac{\sigma\left(I_f\right)}{\mu\left(I_f\right)}. \tag{7}$$

The SC metric takes values from zero (no speckle) to one (fully developed speckle) and is related to the equivalent number of looks by $\mathrm{ENL} = 1/\mathrm{SC}^2$. It assumes that the speckle is invariant across the image. There is a variation of this metric, named measure of contrast (MC) that is applied to the whole filtered image.

The speckle suppression index (SSI) [27] is the ratio of speckle contrast of the filtered image to the speckle contrast of the original noisy image I_n

$$\mathrm{SSI} = \frac{\mathrm{SC}\left(I_f\right)}{\mathrm{SC}\left(I_n\right)}. \tag{8}$$

This metric decreases with increasing filtering strength.

The speckle suppression and mean preservation index (SSMPI) [27] has been developed to circumvent the limitations of ENL and SSI when the filter overestimates the mean value. It is defined as follows

$$\mathrm{SSMPI} = Q\frac{\sigma\left(I_f\right)}{\sigma\left(I_n\right)}, \tag{9}$$

where $Q = 1 + \left|\mu\left(I_f\right) - \mu\left(I_n\right)\right|$. Lower values of this index indicate better performance of the filter in terms of mean preservation and noise reduction.

B. General Quality Metrics

The Laplacian (LAP) is a second order differentiation measure obtained by convolving the filtered image with a Laplacian mask and computing its mean square value

$$\mathrm{LAP} = \sum_{i,j}\Big[I_f\left(i-1,j\right) + I_f\left(i+1,j\right) + I_f\left(i,j-1\right) + I_f\left(i,j+1\right) - 4I_f\left(i,j\right)\Big]^2. \tag{10}$$

Another example of a differentiation metric is the image gradient (GRAD) [33].

Some quality measures evaluate statistical properties of denoised images. The variance (VAR) metric is an example [34]. The autocorrelation (ACR) measure compares all possible pixel pairs and reports the likelihood that both will be bright as a function of the distance and direction of separation. The mathematical definition of ACR for a denoised image I_f is given by

$$\mathrm{ACR}\left(m,n\right) = \sum_{i,j} I_f\left(i,j\right) \times I_f\left(i-m, j-n\right), \tag{11}$$

where $I_f(i,j)$ is the image intensity at pixel (i,j), and m and n represent the distance from the corresponding i and j position.

The signal-to-distortion ratio (SDR) is calculated as the ratio of l_2-norm of the noisy image to the l_2-norm of the noise removed by the filter

$$\mathrm{SDR} = 10\log_{10}\left[\frac{\sum_{i,j} I_n\left(i,j\right)^2}{\sum_{i,j}\left[I_n\left(i,j\right) - I_f\left(i,j\right)\right]^2}\right]. \tag{12}$$

The edge preservation coefficient (EPC) [35], also known as the beta-metric, is a correlation metric that can be used to quantify edge preservation, according to the expression

$$\mathrm{EPC} = \frac{\sum_{i,j}\left(\Delta I_n - \overline{\Delta I_n}\right)\left(\Delta I_f - \overline{\Delta I_f}\right)}{\sqrt{\sum_{i,j}\left(\Delta I_n - \overline{\Delta I_n}\right)^2 \sum_{i,j}\left(\Delta I_f - \overline{\Delta I_f}\right)^2}}, \tag{13}$$

where ΔI_n and ΔI_f are high pass filtered versions of the noisy image and the filtered image, respectively, obtained by an approximation of the Laplacian operator using a 3×3 kernel. The $\overline{\Delta I_n}$ and $\overline{\Delta I_f}$ are the mean values after the operator is applied.

So far, only metrics that have been tested on digital holograms have been described. For completeness, a number of metrics that are commonly used on natural images were added. Some of them take into account the psycho-physical features of the human visual system. The following no-reference metrics were included in this study: the frequency threshold metric (FTM) [36], the Marziliano blurring metric (MARZ) [37], the cumulative probability blur detection (CPBD) [38], the just noticeable blur metric (JNBM) [39], the HP metric (HPM) [40], the kurtosis based metric (KURTZ) [41], the kurtosis of wavelet coefficients (KURTW) [42], the Riemannian tensor based metric (RMN) [43], and the perception based image quality evaluator (PIQE) [44].

C. Evaluation of Objective Metrics

After evaluating the IQMs described above for the tested hologram reconstructions, their values are compared with the available MOS curves, using the statistical relations described hereafter.

Both MOS and IQMs are normalized prior to comparison, for the set of enhanced images of phase 1 given by each filter separately. Then, according to Rec. ITU-R BT.500-13 [29], a logistic model is fitted on the IQMs results to approximate the MOS values and force their relation to be monotonic over its range. The normalized MOS values can be approximated by a logistic function $\mathrm{MOS}_n = f(\mathrm{MR})$ generally defined as

$$\mathrm{MOS}_p = b_1 + \frac{b_2}{1 + \exp\left[-b_3 \times \left(\mathrm{MR} - b_4\right)\right]}, \tag{14}$$

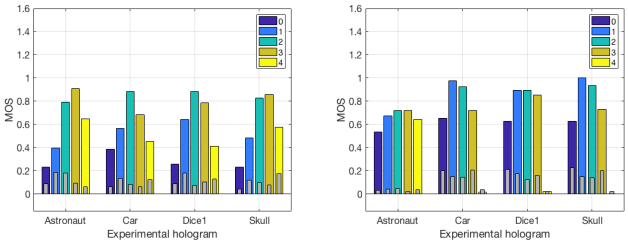
where MR is the objective measure result, MOS_p is the predicted MOS for the given image, and b_i, $i = 1, \ldots, 4$ denote the regression parameters.

To compare subjective results with objective metrics, we use a series of performance indexes. In particular, the Pearson correlation coefficient (PCC), the Spearman rank order correlation coefficient (SROCC), and the root mean squared error (RMSE) are computed, to measure the linearity, monotonicity, and accuracy of the results.

5. RESULTS

A. Subjective evaluation results

The MOS ratings and respective 95% confidence intervals were computed over the set of 16 participants. They are presented in Figure 4 organized per filter parameter index and for the set of 4 tested holograms. It can be seen that the NLM and Lee filters have the highest dynamic range for most objects. This reveals that the degradation of image quality with increasing parameters is perceived more clearly for these filters.



(a) *BM3D* (b) *NLM*

(c) *Lee* (d) *Frost*

Fig. 4. MOS results (colored bars) and corresponding 95% confidence intervals (gray bars) per filter parameter index, for the set of 4 experimental holograms and for the 4 filters.

The parameter index corresponding to the maximum MOS value for each hologram and each filter were selected. These optimal parameter indexes are reported in Table 2. The corresponding parameter values were presented above in Table 1.

Table 2. Optimal filter parameters.

Hologram	Optimal parameter index			
	BM3D	NLM	Lee	Frost
Astronaut	3	2	2	2
Car	2	1	2	2
Dice1	2	2	3	1
Skull	3	1	2	2

The MOS scores and corresponding 95% confidence intervals per filter at the respective optimal parameter value, showing

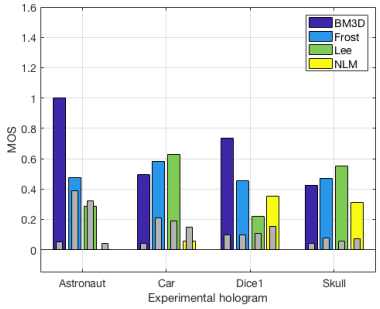


Fig. 5. MOS results (colored bars) and corresponding 95% confidence intervals (gray bars) per filter at the respective optimal parameter value.

the best result for each hologram reconstruction are reported in Figure 5. It can be observed that the BM3D filter gives the highest MOS for the Astronaut and the Dice1 holograms, while the Lee filter assumes the maximum score for the Car and the Skull. A greater contrast between the first and the second positions of the rank can be observed on the MOS plots for the Astronaut and the Dice1 holograms. This indicates a more distinct preference of BM3D filter over the remaining filters, for the referred holograms. Moreover, the MOS plots for the Car and Skull holograms present similar results for BM3D, Frost and Lee.

The Astronaut and the Dice1 holograms don't have many irregular shading transitions or texture variations producing different speckle grain sizes. Hence, a filter that handles speckle removal efficiently while preserving details, such as the BM3D filter, is more likely to succeed in these cases.

The last position in the ranking order is always occupied by the NLM filter. Considering comparative studies based on objective metrics, this behavior is unexpected, since the NLM filter belongs to a class of more recent and advanced filters. The set of parameters used for this method is atypical. However, it was driven by a series of pre-tests that revealed that as soon as filter strength increased for more effective speckle removal, a series of artifacts appeared and the perceived quality decreased quickly. It may be due to the fact that the large size of the displayed images revealed artifacts that usually remain hidden when their natural sizes are depicted. Such artifacts may have no significant impact on objective metrics, but reveal a hindering effect on texture related subjective appearance. A recent comparative study [24] on amplitude holograms, using a USAF 1951 target and objective metrics such as the PSNR or the RMSE gave a ranking order of BM3D, NLM, Frost and finally the Lee filter. The fact that such test object has no texture or shading effects, nor even depth variations, may explain the observed differences. An earlier study by the same authors on phase images had a quite different outcome since the top of the rank was occupied by different filters such as the 2D windowed Fourier transform filter [45], the curvelets based filter [46], or the SPADEDH method [47], and BM3D occupied at most the third position. It must be stressed that not only the nature of the holograms may explain the absence of similar results in the literature, but specially the distinct nature of subjective assessment.

B. Objective evaluation results

The absolute values of Pearson and Spearman rank order correlation coefficients for all studied objective metrics are depicted

in Figures 6 and 7, respectively. The correlations have been separated by filter, since it was found that their different behaviours affect IQMs performance and result in poorer outcomes when they are analyzed altogether. It can be easily noted that the correlations are much lower when considering the BM3D method, while for the NLM (the least performing filter, from the perceptual point of view) the correlations are much higher. For the Lee and the Frost filters, the performance is more identical, with a slight advantage when IQMs are applied to the Lee enhanced images.

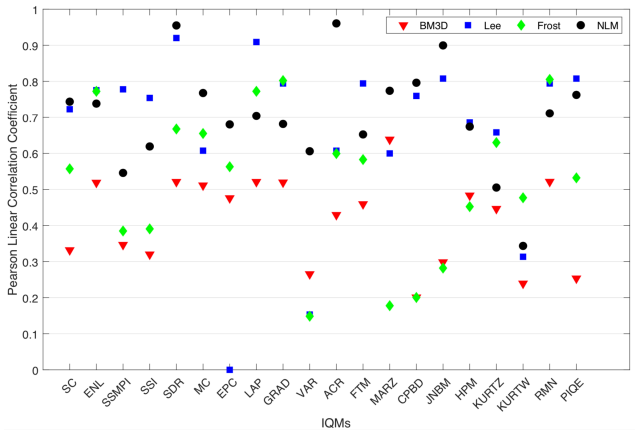


Fig. 6. Absolute values of Pearson linear correlation coefficient for the studied IQMs.

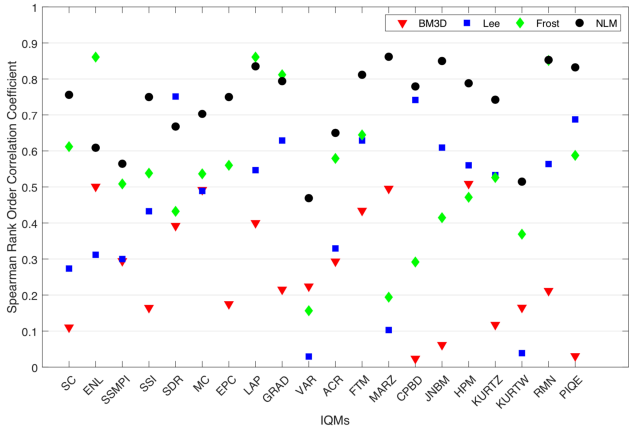


Fig. 7. Absolute values of Spearman rank order correlation coefficient for the studied IQMs.

The results obtained for the 20 IQMs allowed to identify the best performing metrics. In this sense, the two IQMs with higher SROCC were selected for each filter for further analysis. Considering the 4 filters, this amounts to a total of 8 metrics. The performance measures for these metrics are presented in Table 3 and the logistic fits obtained to approximate the MOS curves are depicted in Figure 8.

For the BM3D filter, it can be observed that the metrics perform very poorly. This can be due to the complex nature of this advanced filter that doesn't limit its function to smoothing the grainy appearance of speckle, but also performs a number of

Table 3. Performance measures of the selected IQMs.

IQM	SROCC	PCC	RMSE
HPM (BM3D)	0.509	0.483	0.298
ENL (BM3D)	0.501	0.519	0.291
SDR (Lee)	0.751	0.920	0.116
CPBD (Lee)	0.742	0.760	0.195
ENL (Frost)	0.861	0.772	0.153
LAP (Frost)	0.861	0.772	0.153
MARZ (NLM)	0.862	0.774	0.217
RMN (NLM)	0.853	0.711	0.235

image enhancement tasks. Since different images present diverse characteristics, the filter adapts in different ways and this leads to a distinct IQMs behaviour for each hologram. Filters with a simpler statistical behaviour, that are specifically tailored for speckle noise but have less edge preservation and image enhancement capabilities, such as the Lee and the Frost algorithms, lead to more consistent behaviour across different images. The NLM filter presents higher correlations, but this could be due to the fact that the observers preferred images with higher speckle contents when compared to the other filters since the artifacts introduced by the NLM could be very annoying at the displayed resolution. Thus, for the range of parameters tested, the function of this filter was mainly to smooth the speckled zones. At this stage, blurring was not a disturbing distortion yet, that would impact the metrics performance, but several artifacts could al-

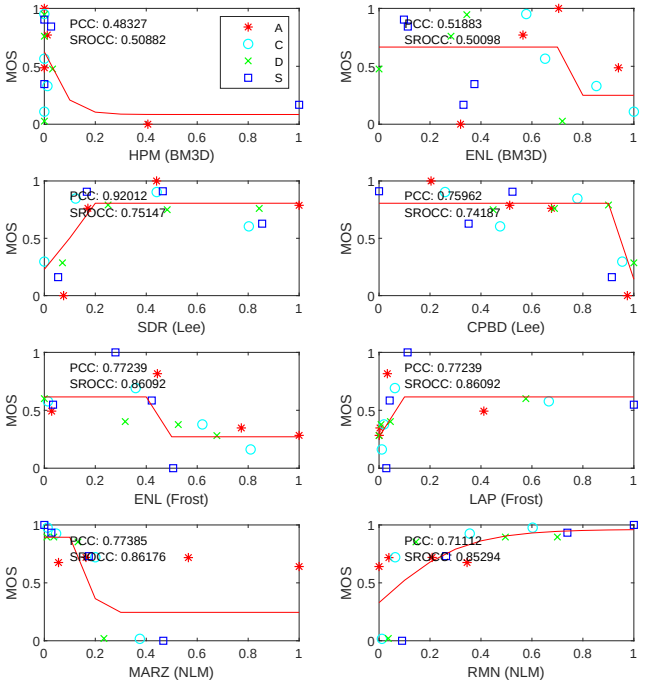


Fig. 8. Selected IQMs fits corresponding to Table 3. Each plot displays the results of the four objects, namely the Astronaut (A), Car (C), Dice1 (D) and Skull (S).

ready be noticed by the observers.

In spite of the higher values of the correlation coefficients in the NLM case, it is suggested by Figure 8 that for the Frost and the Lee filters, the IMQs are more reliable and may present an easier way to estimate the optimal filter parameter. One of the main difficulties for comparing IQMs and MOS is related to the non-monotonous behaviour of the latter measure. However, if a relation between the saturation point of a quality metrics curve and the optimal filter parameter obtained by MOS could be found, an acceptable objective estimate would be possible.

6. CONCLUSIONS

The subjective and objective quality assessment of 4 speckle denoising filters, namely the NLM, the Lee, the Frost and the BM3D filters, using 4 parameter settings in each case, was performed. The subjective results were ranked with respect to the perceived image quality to obtain the MOS for each filter/parameter combination. After obtaining the best combination for each filter, they were subjectively compared once again to obtain the MOS based ranking of the set of 4 filters.

To compare the subjective outcomes with objective measures, 20 no-reference image quality metrics were evaluated for the set of 16 holographic images corresponding to different holograms and filter/parameter combinations. The comparison was performed using statistical indexes such as the PCC and the SROCC. The experiment indicates that BM3D and Lee are the preferred filters, and reveals a strong dependence of filter performance on the hologram characteristics. The BM3D method clearly yielded the best results for the Astronaut and Dice1 holograms, followed by the Frost filter. The Lee filter was the preferred filter for the Car and the Skull, followed closely by the the Frost and the BM3D filters. When comparing the MOS results with the image quality metrics, the best performing measures were the HPM for the BM3D filter, the SDR for the Lee filter, the ENL for the Frost filter, and the MARZ for NLM filter.

The subjective results of this work reveal that there is a strong dependence of filter performance on the hologram characteristics. Furthermore, the comparison of objective quality metrics with the MOS indicate different behaviour for different filters, and reveal trends for future development of hologram oriented image quality metrics.

Funding. Portuguese FCT and FEDER PT2020 partnership agreement under the projects PTDC/EEI-PRO/2849/2014 - POCI-01-0145-FEDER-016693, and UID/EEA/50008/2019.

REFERENCES

1. V. Bianco, P. Memmolo, M. Leo, S. Montresor, C. Distante, M. Paturzo, P. Picart, B. Javidi, and P. Ferraro, "Strategies for reducing speckle noise in digital holography," Light. Sci. & Appl. **7**, 48 (2018).
2. T. Nomura, M. Okamura, E. Nitanai, and T. Numata, "Image quality improvement of digital holography by superposition of reconstructed images obtained by multiple wavelengths." Appl. Opt. **47**, D38–D43 (2008).
3. Y. Park, W. Choi, Z. Yaqoob, R. Dasari, K. Badizadegan, and M. S. Feld, "Speckle-field digital holographic microscopy." Opt. express **17**, 12285–12292 (2009).
4. K. M. Molony, J. Maycock, J. B. McDonald, B. M. Hennelly, and T. J. Naughton, "A Comparison of wavelet analysis techniques in digital holograms," in *Photon Management III,* , vol. 6994 (2008), pp. 699412–699412–8.
5. R. Srivastava, J. R. Gupta, and H. Parthasarthy, "Comparison of PDE based and other techniques for speckle reduction from digitally reconstructed holographic images," Opt. Lasers Eng. **48**, 626–635 (2010).
6. S. Montresor and P. Picart, "Quantitative appraisal for noise reduction in digital holographic phase imaging," Opt. Express **24**, 14322 (2016).
7. T. M. Lehtimäki, K. Sääskilahti, and T. J. Naughton, "Visual perception of digital holograms on autostereoscopic displays," in *Proc. SPIE 7329, Three-Dimensional Imaging, Visualization, and Display,* , vol. 7329 (2009), pp. 1–7.
8. T. M. Lehtimäki, K. Sääskilahti, T. Pitkäaho, and T. J. Naughton, "Evaluation of perceived quality attributes of digital holograms viewed with a stereoscopic display," in *2010 9th Euro-American Workshop on Information Optics,* (2010), pp. 1–3.
9. T. M. Lehtimäki, K. Sääskilahti, T. Pitkäaho, and T. J. Naughton, "Comparing numerical error and visual quality in reconstructions from compressed digital holograms," in *SPIE Defense, Security, and Sensing,* , vol. 7690 (2010), p. 769012.
10. E. Darakis, M. Kowiel, R. Näsänen, and T. J. Naughton, "Visually lossless compression of digital hologram sequences," in *Image Quality and System Performance VII,* , vol. 7529 (International Society for Optics and Photonics, 2010), p. 752912.
11. A. Ahar, D. Blinder, T. Bruylants, C. Schretter, A. Munteanu, and P. Schelkens, "Subjective quality assessment of numerically reconstructed compressed holograms," in *Applications of Digital Image Processing XXXVIII,* , vol. 9599 (International Society for Optics and Photonics, 2015), p. 95990K.
12. A. Ahar, T. Birnbaum, C. Jäh, and P. Schelkens, "A new similarity measure for complex amplitude holographic data," in *Applications of Digital Image Processing XL*, (2017), p. 54.
13. A. Ahar, T. Birnbaum, D. Blinder, A. Symeonidou, and P. Schelkens, "Performance evaluation of sparseness significance ranking measure (ssrm) on holographic content," in *Imaging and Applied Optics 2018 (3D, AO, AIO, COSI, DH, IS, LACSEA, LS&C, MATH, pcAOP),* (Optical Society of America, 2018), p. JTu4A.10.
14. T. M. Lehtimäki, R. G. Reilly, and T. J. Naughton, "Towards perception-inspired numerical measures of compression error in digital holograms of natural three-dimensional scenes," in *Imaging and Applied Optics 2018,* (Optical Society of America, 2018), p. JTu4A.42.
15. E. S. Fonseca, P. T. Fiadeiro, M. Pereira, and A. Pinheiro, "Comparative analysis of autofocus functions in digital in-line phase-shifting holography," Appl. optics **55**, 7663–7674 (2016).
16. M. V. Bernardo, P. Fernandes, A. Arrifano, M. Antonini, E. Fonseca, P. T. Fiadeiro, A. M. Pinheiro, and M. Pereira, "Holographic representation: Hologram plane vs. object plane," Signal Process. Image Commun. **68**, 193–206 (2018).
17. J. Dainty, *Laser Speckle and Related Phenomena*, Topics in Applied Physics (Springer Berlin Heidelberg, 2013).
18. J. W. Goodman, *Speckle phenomena in optics: theory and applications* (Roberts and Company Publishers, 2007).
19. K. Dabov and A. Foi, "Image Denoising with Block-matching and {3D} Filtering," Electron. Imaging **6064**, 1–12 (2006).
20. K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Image denoising by sparse 3-d transform-domain collaborative filtering," IEEE Transactions on Image Process. **16**, 2080–2095 (2007).
21. A. Buades, B. Coll, and J. . Morel, "A non-local algorithm for image denoising," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05),* , vol. 2 (2005), pp. 60–65 vol. 2.
22. J. S. Lee, "Digital image enhancement and noise filtering by use of local statistics," IEEE Transactions on Pattern Analysis Mach. Intell. **PAMI-2**, 165–168 (1980).
23. V. S. Frost, J. A. Stiles, K. S. Shanmugan, and J. C. Holtzman, "A model for radar images and its application to adaptive digital filtering of multiplicative noise," IEEE Transactions on Pattern Analysis Mach. Intell. **PAMI-4**, 157–166 (1982).
24. S. Montresor and P. Picart, "Evaluation of de-noising algorithms for amplitude image restoration in digital holography," in *Applied Industrial Optics: Spectroscopy, Imaging and Metrology,* (Opt. Soc. of America, 2018), pp. JTu4A–8.
25. S. Montresor and P. Picart, "Quantitative appraisal for noise reduction in digital holographic phase imaging," Opt. Express **24**, 14322–14343

(2016).
26. P. Coupe, P. Hellier, C. Kervrann, and C. Barillot, "Nonlocal means-based speckle filtering for ultrasound images," IEEE Transactions on Image Process. **18**, 2221–2229 (2009).
27. A. Uzan, Y. Rivenson, and A. Stern, "Speckle denoising in digital holography by nonlocal means filtering." Appl. optics **52**, A195–200 (2013).
28. L. L. Thurstone, "A law of comparative judgment," Psychol. Rev. **34**, 273–286 (1927).
29. ITU-R Recommendation BT.500-13, "Methodology for the subjective assessment of the quality of television pictures," Tech. rep., International Telecommunication Union (2012).
30. ITU-R Recommendation BT.2022, "General viewing conditions for subjective assessment of quality of sdtv and hdtv television pictures on flat panel displays," Tech. rep., International Telecommunication Union (2012).
31. R. A. Bradley and M. E. Terry, "Rank analysis of incomplete block designs: I. the method of paired comparisons," Biometrika. **39**, 324–345 (1952).
32. K. Tsukida and M. R. Gupta, "How to analyze paired comparison data," Tech. rep., Department of Electrical Engineering, University of Washington, Seattle, Washington, USA (2011).
33. L. Liu, Y. Hua, Q. Zhao, H. Huang, and A. C. Bovik, "Blind image quality assessment by relative gradient statistics and adaboosting neural network," Signal Process. Image Commun. **40**, 1–15 (2016).
34. S. Erasmus and K. Smith, "An automatic focusing and astigmatism correction system for the sem and ctem," J. Microsc. **127**, 185–199 (1982).
35. J. L. Mateo and A. Fernández-Caballero, "Finding out general tendencies in speckle noise reduction in ultrasound images," Expert. Syst. with Appl. **36**, 7786–7797 (2009).
36. A. V. Murthy and L. J. Karam, "A matlab-based framework for image and video quality evaluation," in *2010 Second International Workshop on Quality of Multimedia Experience (QoMEX),* (IEEE, 2010), pp. 242–247.
37. P. Marziliano, F. Dufaux, S. Winkler, and T. Ebrahimi, "Perceptual blur and ringing metrics: application to jpeg2000," Signal processing: Image communication **19**, 163–172 (2004).
38. N. D. Narvekar and L. J. Karam, "A no-reference image blur metric based on the cumulative probability of blur detection (cpbd)," IEEE Transactions on Image Process. **20**, 2678–2683 (2011).
39. R. Ferzli and L. J. Karam, "A no-reference objective image sharpness metric based on the notion of just noticeable blur (jnb)," IEEE transactions on image processing **18**, 717–728 (2009).
40. D. Shaked and I. Tastl, "Sharpness measure: Towards automatic image enhancement," in *IEEE International Conference on Image Processing 2005,* , vol. 1 (IEEE, 2005), pp. I–937.
41. N. F. Zhang, A. Vladar, M. T. Postek, and R. D. Larrabee, "A kurtosis-based statistical measure for two-dimensional processes and its applications to image sharpness," Tech. rep. (2003).
42. R. Ferzli, L. J. Karam, and J. Caviedes, "A robust image sharpness metric based on kurtosis measurement of wavelet coefficients," in *Proc. of Int. Workshop on Video Processing and Quality Metrics for Consumer Electronics,* , vol. 12 (2005).
43. R. Ferzli and L. J. Karam, "A no reference objective shaprness metric using riemannian tensor," Simulation **1**, 1 (2007).
44. Venkatanath N, Praneeth D, Maruthi Chandrasekhar Bh, S. S. Channappayya, and S. S. Medasani, "Blind image quality evaluation using perception based features," in *2015 Twenty First National Conference on Communications (NCC),* (2015), pp. 1–6.
45. Q. Kemao, "Windowed fourier transform for fringe pattern analysis," Appl. Opt. **43**, 2695–2702 (2004).
46. J.-L. Starck, E. Candès, and D. Donoho, "The curvelet transform for image denoising," IEEE Trans. image processing **11**, 670–684 (2002).
47. P. Memmolo, M. Iannone, M. Ventre, P. Netti, A. Finizio, M. Paturzo, and P. Ferraro, "Quantitative phase maps denoising of long holographic sequences by using spadedh algorithm," Appl. optics **52**, 1453–1460 (2013).