A Vector Approach for Image Quality Assessment and Some Metrological Considerations

Alessio De Angelis, *Student Member, IEEE*, Antonio Moschitta, *Member, IEEE*, Fabrizio Russo, *Senior Member, IEEE*, and Paolo Carbone, *Associate Member, IEEE*

***Abstract*—The main purpose of this paper is to present a metrology-based view of the image quality assessment (IQA) field. Three main topics are developed. First, the state of the art in the field of IQA is presented, providing a classification of some of the most important objective and subjective IQA methods. Then, some aspects of the field are analyzed from a metrological point of view, also through a comparison with the software quality measurement area. In particular, a statistical approach to the evaluation of the uncertainty for IQA objective methods is presented, and the topic of measurement modeling for subjective IQA methods is analyzed. Finally, a vector approach for full-reference IQA is discussed, with applications to images corrupted by impulse and Gaussian noise. For these applications, the vector root mean squared error (VRMSE) and fuzzy VRMSE are introduced. These vector parameters provide a possible way to overcome the main limitations of the mean-squared-error-based IQA methods.**

***Index Terms*—Fuzzy methods, fuzzy sets, Gaussian noise, image analysis, image quality assessment (IQA), impulse noise, Q factor, vector root mean squared error (VRMSE).**

I. Introduction

IMAGE quality assessment (IQA) plays an important role in every aspect of the digital image processing field. Its applications range from acquisition devices to compression and communication systems. It can be used as a feedback tool, to monitor visual quality, optimize parameters and algorithms, or benchmark processing systems [1]. It is of considerable interest also for image-based instrumentation, where it is used to validate algorithms of fundamental importance such as image enhancement and noise cancellation techniques and to provide the automatic tuning of their parameters, as requested in many real applications [2].

After a description of the state of the art in the IQA field, the primary purpose of this paper is to present a metrological characterization of the various IQA methods, focusing on quality measurement uncertainty. Furthermore, this metrological approach is also discussed through an application to objective full-reference (FR) IQA for the performance assessment of noise-reduction filters. For this application, a vector approach is proposed, and the vector root mean squared error (VRMSE) parameter is presented. By introducing the fuzzy VRMSE, this paper extends the results presented in [3] to comprehend the case of images corrupted by Gaussian noise. Furthermore, the vector approach is also extended to color images.

This paper is organized as follows: Section II delineates an overview of the state of the art. Section III discusses the statistical evaluation of the performance for objective IQA methods. Section IV focuses on the metrological characterization. Section V presents the vector approach, and Section VI reports conclusions.

Manuscript received September 16, 2007; revised March 5, 2008. First published September 19, 2008; current version published December 9, 2008. The Associate Editor coordinating the review process for this paper was Prof. Alessandro Ferrero.

A. De Angelis, A. Moschitta, and P. Carbone are with the Department of Electronic and Information Engineering, University of Perugia, 06125 Perugia, Italy (e-mail: deangelis@diei.unipg.it; moschitta@diei.unipg.it; carbone@diei.unipg.it).

F. Russo is with the Dipartimento di Elettrotecnica, Elettronica e Informatica, University of Trieste, 34127 Trieste, Italy (e-mail: rusfab@univ.trieste.it).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIM.2008.2004982

II. State of the Art

The IQA problem can be addressed in two complementary ways: subjective and objective methods. The subjective IQA methods consist of the collection and processing of opinions from observers by means of statistical techniques, whereas the objective methods aim at producing a quantitative measure that can automatically predict the perceived image quality. The two classes of methods are strictly connected, because the perceived quality predictions provided by the objective methods can be validated through comparison with the results of the subjective evaluations. Furthermore, the performance of the objective methods depends on the degree of correlation with the subjective methods.

In application areas where the image quality is an intrinsically subjective concept, the most logical and accurate way to evaluate it is to solicit the subjective opinions of observers. Furthermore, the perception of quality may be influenced by many factors such as context, degree of interaction, and psychological factors such as visual attention and recency effect [4]. A standardized procedure for subjective IQA called the double stimulus continuous quality scale is described in the ITU-T Recommendation BT.500-11 [5]. In this procedure, the observers are presented with couples of images (reference and distorted), the order of which is randomized during the test session. The subjects rate each image on a continuous quality scale. At the end of the test session, the quality scores are linearly mapped to obtain raw scores and subsequently normalized and processed to obtain the mean opinion scores (MOSs). A measure of the perceived relative quality of the distorted image with respect to the reference is then given by the difference MOS (DMOS).

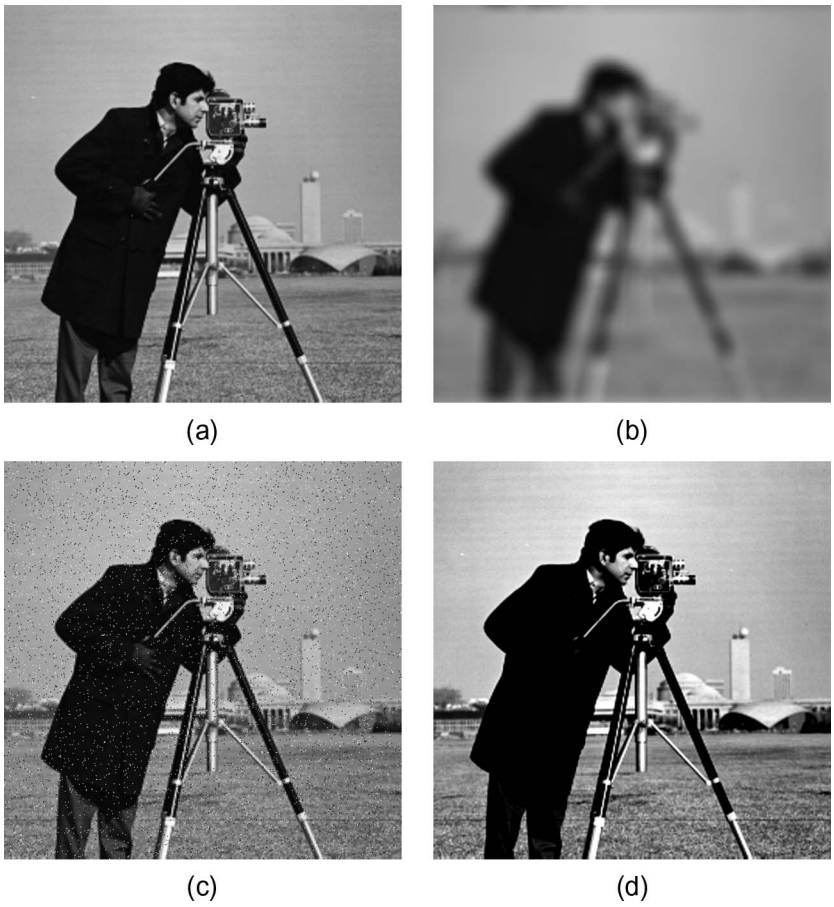


Fig. 1. "Cameraman" image. (a) Original. (b) Blurred. (c) Corrupted with "salt and pepper" noise. (d) Contrast stretched.

These types of procedures, however, are time consuming and not feasible in most practical applications. Thus, there is the need to develop efficient objective IQA methods. Most of the objective methods for IQA can be defined as either FR or no reference (NR). The FR methods require the comparison between an original image, which is assumed to have perfect quality, and a degraded version of the same image. These methods are widely used for testing and validating image-based systems and applications. Conversely, the NR methods provide an estimate of quality without any knowledge of the original image. This allows them to be implemented in algorithms, but usually with limited scope. They are designed to estimate visual quality only in images corrupted with specific distortion types, such as blocking (artifact of Joint Photographers Expert Group (JPEG) compression) [6], [7], blurring, and ringing (JPEG 2000) [8]. Recently, another class of methods has been introduced, i.e., reduced reference, in which the IQA algorithm has access to some information about the reference in a side channel or embedded in the degraded image [9].

The most commonly used FR image quality measurements are based on the computation of pixelwise differences between the original and the distorted image, such as mean squared error (MSE) and peak signal-to-noise ratio (PSNR). These methods have the advantage of low computational complexity and ease of implementation, but it is well known that they are not well correlated with the results of the subjective methods. To illustrate this behavior, Fig. 1 shows an original grayscale image, i.e., the 512×512 "Cameraman" test picture (8 bit/pixel), along with three distorted versions. For all the distorted images, the PSNR is 20.04 dB, but their perceived quality is clearly not the same. For this reason, the goal of the research in this field has been to find more accurate measurement strategies. In the literature, numerous examples can be found, which can be divided into various groups. Apart from PSNR, some other notable methods in the mathematical distance formulations group are based on the spectral distance of images, the local histogram [10], and soft computing approaches [11]. Another group takes into account error sensitivity by modeling the behavior of the human visual system (HVS), based on neurophysiological research and psychophysical experiments on the threshold of visual perception [12]. Furthermore, a different approach has recently been proposed, i.e., the so-called structural approach, which led to the structural similarity index [13]. It quantifies perceived distortion by measuring the loss of structural information. An additional group of methods is based on the information-theoretic approach [14], [15]. In particular, the visual information fidelity measure [14], by combining the mutual information between the input and the output of the HVS channel model in both cases of the presence and absence of distortion, quantifies the information in the reference image and the loss of information due to the channel. Other approaches include feature-based modeling [16] and HVS with application-specific modeling [17].

III. Statistical Evaluation

To evaluate the performance of the various objective IQA methods, a possible way is to compare their results with those provided by the subjective methods. An example of a performance evaluation process is the procedure employed in the Video Quality Experts Group Phase I FR-TV test [18]. This procedure has been specifically designed for video quality testing, but it has also recently been applied to the quality assessment of still images [19]. Various IQA algorithms are extensively tested over a wide range of original images and distortion types. The performance is analyzed according to three main criteria. The first criterion is prediction *accuracy*. This is the ability of the objective algorithm to predict the subjective quality ratings with low error, which is evaluated by means of the Pearson linear correlation coefficient between the subjective and objective scores after nonlinear regression. The second criterion is prediction *monotonicity*, which is the degree to which the model's predictions agree with the relative magnitudes of the subjective quality ratings. This is analyzed using the Spearman rank order correlation coefficient. Finally, the third criterion is prediction *consistency*, which is the degree to which the model maintains prediction accuracy over the range of test images and distortion types, which are measured by means of the outlier ratio [13], [18]. To establish whether the differences between the performance of methods are statistically significant, a hypothesis test is conducted by means of techniques such as the analysis of variance.

The results of the recent statistical FR IQA performance evaluation study [19] have shown that none of the considered methods is statistically superior over all of the others in every type of distortion. It is also possible to note that the PSNR metric has the best performance in the presence of white Gaussian noise, whereas, for the other classes of distortion considered in the study (blur, JPEG 2000 compression, and simulated fast-fading Rayleigh wireless channel), other methods have considerably better behavior.

Another approach for the validation of objective methods using subjective assessment is presented in [20], for the specific application field of image fusion algorithms. This approach uses several subjective performance metrics, i.e., the total preference score, scene preference score, and individual preference score. Furthermore, two methods for subjective–objective correspondence evaluation are proposed: 1) the correct ranking measure and 2) the subjective relative measure.

IV. Metrological Characterization

The main characteristic of the majority of the objective FR IQA methods presented in Section II is that they aim to provide a universal quality measure that is a metric that correlates well with the subjective ratings in all the possible forms of image degradation. However, as pointed out in Section III, for different distortion types, there are differences in the behavior of the proposed methods. Furthermore, there is no single quality measure that gives better results than the others for all types of distortion.

Given the difficulties of examining the IQA problem, a comparison could be made with other areas that face similar issues, such as the software quality assessment field, in which the definition of product quality is complex. In this field, the ISO/IEC 9126-1:2001 International Standard [21] provides a quality model that classifies software quality in a structured set of characteristics, i.e., functionality, reliability, usability, efficiency, maintainability, and portability. These main characteristics and their subcharacteristics are associated with attributes, which are entities that can be verified or measured in the software product. Moreover, for each attribute, it is possible to define one or more specific metrics. A similarity could be established between these characteristics and attributes and the various aspects that compose the IQA problem. As an example, single distortion types affecting an image (such as additive noise, blurring, ringing, or blocking) could be considered as attributes for quality evaluation. Consequently, specific metrics for each attribute could be designed. It is possible to note that two of the main properties that these metrics should have are *selectivity* and *validity*. *Selectivity* is related to the ability to provide the measurement results of a particular attribute that are not influenced by the other attributes. *Validity* can be related to "whether the measurement result provides sufficient information for support decision that is useful to the measurement purpose" [22], [23].

IQA methods can be analyzed according to the principles of the Guide to the expression of Uncertainty in Measurement (GUM) [24]. The uncertainty associated with the result of an objective quality measurement can be evaluated by applying a statistical approach, thus using a type-A evaluation method. Following this approach, the uncertainty can be evaluated by calculating the standard deviation of the quality prediction generated by the IQA system for each image on which the experiment is executed. To provide an example of such an approach, a publicly available database was used [25]. The database contains 779 test images used in an extensive subjective quality assessment study [19], together with the relative DMOS scores. The test images are highly diversified and are corrupted with various types and levels of degradations. The degraded images in the database have initially been grouped into several subsets identified by a range of DMOS values. Subsequently, for each image in a subset, the PSNR has been calculated with respect to the original image. Fig. 2 shows the mean and standard deviation of the PSNR for each subset. In this case, the DMOS data without realignment are considered, but similar results can be obtained, considering realigned data. The standard deviation of the PSNR, which is expressed in decibels, has been obtained by means of the law of uncertainty propagation. These results, using a type-A evaluation method, lead to a PSNR uncertainty of approximately 6 dB. Fig. 3 shows the results of the application of the same procedure using the Multiscale Structural SIMilarity (MSSIM) index [13] for each subset.

Furthermore, following the approach for input–output measurement modeling used in [26], we can evaluate the various contributions to a subjective image quality measurement result according to

$$Q_{\mathrm{subj}} = Q_{\mathrm{meas}} + \Delta_m + \Delta_q + \Delta_p$$

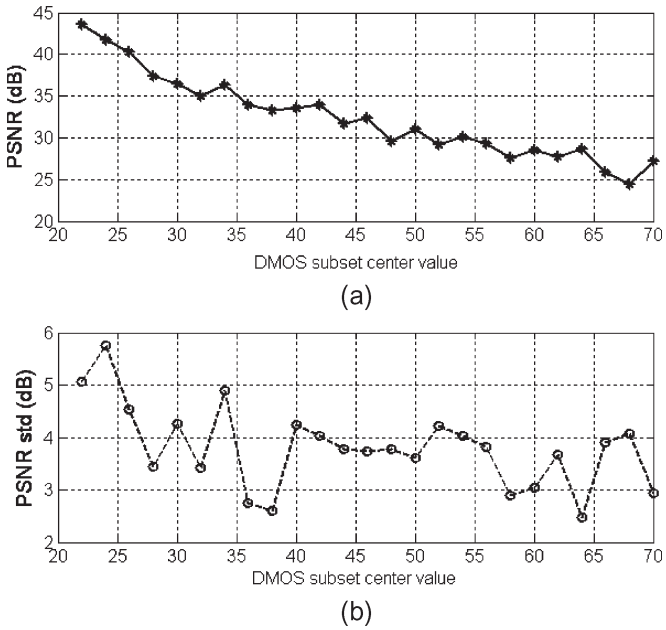


Fig. 2. (a) Mean and (b) standard deviation of the PSNR versus DMOS.

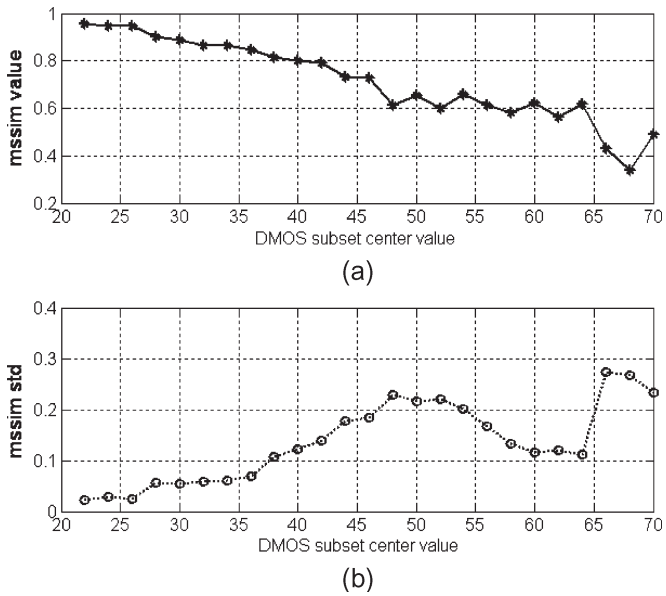


Fig. 3. (a) Mean and (b) standard deviation of the MSSIM versus DMOS.

where

$Q_{\rm subj}$ corrected value of subjective image quality;

$Q_{\rm meas}$ measured subjective image quality, which is the average of a number of opinion scores from human observers (MOS or DMOS);

Δ_m subjective method correction, taking into account the quality deviation due to subjective method imperfections, such as those reported in the following list.

1) Recency effect: Some studies reported that, in subjective IQA sessions, the psychological effect of recency is not negligible [27]. This is particularly important in the quality assessment of video sequences, but such psychological effects also play a role for images [4].
2) Fatigue effect: This effect limits the number of images on which a subjective quality assessment may be conducted. To reduce observer fatigue, the duration of a subjective IQA session should be less than the recommended maximum of 30 min [5], [19].
3) Degree of interaction of the observer with the image: An observer's rating of an image may be different if a level of interaction is required by an application (e.g., surveillance) [4].
4) Degree of expertise of the observers in the field of digital image processing [28].
5) Observers' training [29].

Δ_q image quality correction, which is needed to correct deviations of image quality due to image content.

1) Visual attention: The presence or absence of features and regions of interest in an image, together with the localization of impairments within the picture, may affect a subjective rating of the image.
2) Visual masking effects: The visibility of a stimulus may be decreased due to the presence of another stimulus in the background or the vicinity [12]. As a result, some features, patterns, or artifacts that are present in an image may influence an observer's evaluation of image quality. For example, an image compression artifact or degradation may be masked by image content, so that it falls under the visual detection threshold and is not noticed by the observer.

Δ_p measurement process correction, which is needed to correct deviations of image quality due to imperfections of the measurement process.

1) Differences in the type, setup, and calibration of the monitors.
2) Differences in the environment illumination level.
3) Differences in the viewing distance.
4) Scale mismatch in different sessions: In recent studies, the measurement process is organized in multiple sessions to evaluate a large number of images without observer fatigue effects [19]. In each session, the quality distribution of images is different from the others, and it is not known *a priori*. Thus, there is the need to conduct an *a posteriori* subjective scale realignment session.

V. Vector Approach to IQA

As mentioned in the previous sections, FR image quality measurements such as the MSE (or PSNR) are commonly adopted to validate image-filtering techniques. It is known, however, that the MSE by itself cannot characterize the behavior of a filter with respect to two key features: 1) noise cancellation and 2) detail preservation. Thus, visual inspection is often necessary to take into account these important aspects. As an example, let us consider a 256 $\times$ 256 version of the "Boats" picture [Fig. 4(a)]. Fig. 4(b) shows a noisy version of this image, which was generated by superimposing "salt and pepper" impulse noise with a probability of 24.7%. The results of the application of the 3 $\times$ 3 and 5 $\times$ 5 median filters are shown in Fig. 4(c) and (d), respectively. The processed images

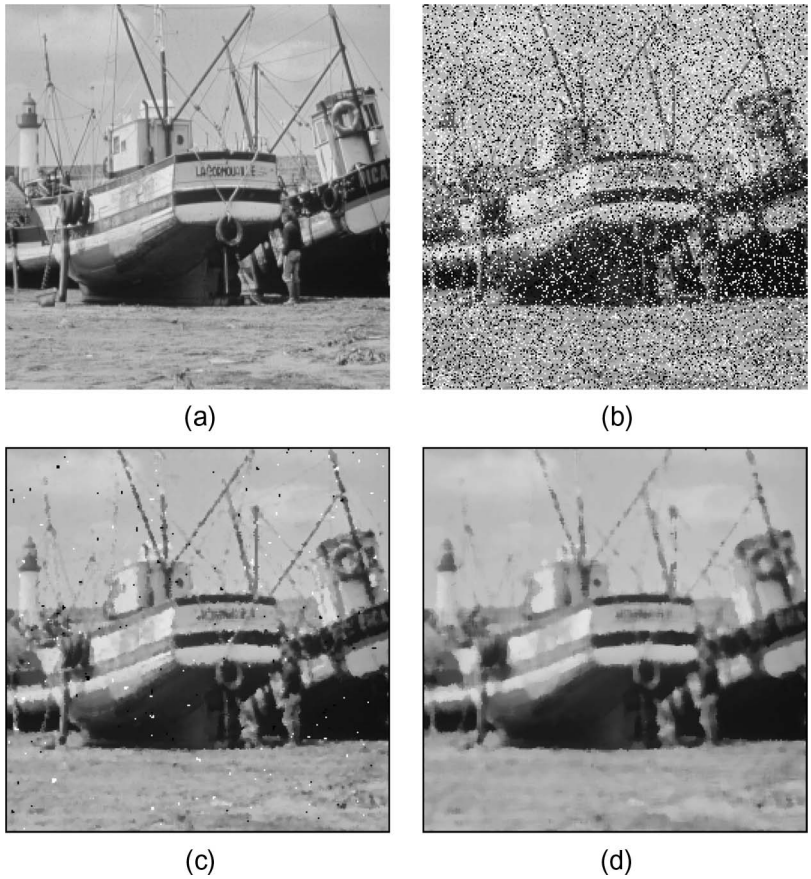


Fig. 4. (a) Original "Boats" picture. (b) Image corrupted by "salt and pepper" impulse noise. (c) Result yielded by the 3×3 median filter. (d) Result yielded by the 5×5 median filter.

have the same MSE $= 272$; however, their visual quality is quite different. Indeed, the 5×5 operator removes all noise pulses at the price of a very annoying image blur [see Fig. 4(d)]. Conversely, the 3×3 median is better at preserving the image details. However, its smoothing performance is much lower, and many noise pulses are still present after filtering [Fig. 4(c)]. Of course, our goal was not to assess the performance of classical median filters, which have extensively been studied in the literature. We have just exploited the well-known effects of different window sizes to obtain an example where two output pictures have the same MSE but different qualities from the point of view of noise cancellation and detail preservation. Clearly, an image quality measurement that aims at replacing visual inspection should be able to perform a separate evaluation of the aforementioned features. A simple idea is to resort to a *vector* approach.

A. Vector Approach for Images Corrupted by Impulse Noise

Formally, suppose that we deal with digitized images having L gray levels (typically, $L = 256$). Let $x(\mathbf{n})$ be the pixel luminance at location $\mathbf{n} = [n1, n2]$ in the image corrupted by impulse noise. Let $y(\mathbf{n})$ and $r(\mathbf{n})$ be the pixel luminances at the same location in the filtered and reference images, respectively. Let S be a region of interest (possibly the entire picture) where we want to evaluate the image quality. Let us introduce the VRMSE as follows:

$$\mathbf{VRMSE} = [\mathrm{RMSE}_A, \mathrm{RMSE}_B] \tag{1}$$

$$\mathrm{RMSE}_A = \sqrt{\frac{1}{N}\sum_{n\in A}[y(\mathbf{n}) - r(\mathbf{n})]^2} \tag{2}$$

$$\mathrm{RMSE}_B = \sqrt{\frac{1}{N}\sum_{n\in B}[y(\mathbf{n}) - r(\mathbf{n})]^2} \tag{3}$$

where N is the total number of processed pixels in region S, $A(\subset S)$ is the subset of coordinates that denotes the group of N_A pixels corrupted by impulse noise $[x(\mathbf{n}) \neq r(\mathbf{n})]$, and $B(\subset S)$ is the subset of coordinates that denotes the group of N_B uncorrupted pixels in the noisy image $[x(\mathbf{n}) = r(\mathbf{n})]$. Since $S = A \cup B$ and $A \cap B = \emptyset$, we have $N = N_A + N_B$. According to (2) and (3), the overall MSE is given by

$$\mathrm{MSE} = \mathrm{RMSE}_A^2 + \mathrm{RMSE}_B^2. \tag{4}$$

Clearly, we have

$$|\mathbf{VRMSE}| = \sqrt{\mathrm{MSE}}. \tag{5}$$

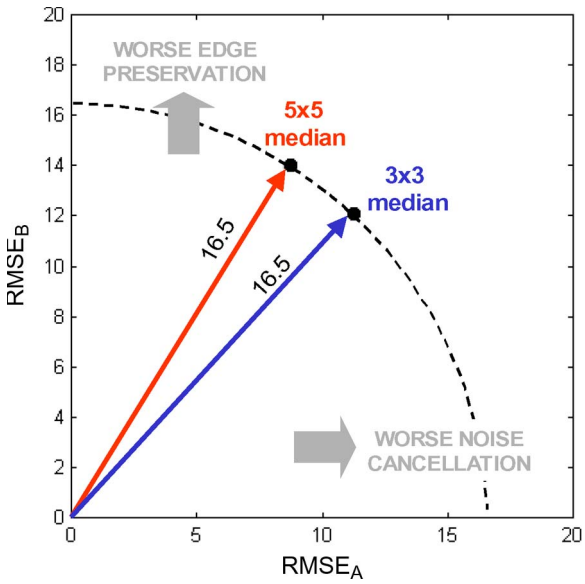


Fig. 5. Example of the graphical representation of the vector errors for the 3×3 and 5×5 median filters ("Boats" picture corrupted by "salt and pepper" impulse noise with a probability of 24.7 %).

The number N_A depends on the amount of noise corruption. In computer simulations, where we superimpose impulse noise with a given probability p to the image, it is expected that $N_A \simeq pN$ and $N_B \simeq (1-p)N$. Fig. 5 shows the evaluation of the vector errors for the filtered images in the previous example. Since the 3×3 and 5×5 medians operate only in the interiors 254×254 and 252×252, respectively, of the input image measuring 256×256, we chose the smaller interior (where both filters operate) to represent region S and $N = 252 \times 252$. According to (5), both vectors in Fig. 5 have the same length ($|\text{VRMSE}| = 16.5$), because the processed pictures [Fig. 4(c) and (d)] have the same $\text{MSE} = 272$. However, the different orientations of these vectors can immediately characterize the well-known behaviors of the filters. Worse detail preservation of the 5×5 median is highlighted by a vector error closer to the vertical axis. Conversely, the worse noise cancellation of the 3×3 operator is denoted by a vector closer to the horizontal axis (Fig. 5). RMSE_A highlights worse noise cancellation, because it is evaluated only in subset A of pixels that have been corrupted by noise. RMSE_A takes into account only the filtering errors that are related to unsatisfactory correction of noise, so its value is zero only in the case of perfect removal of all the noise pulses. On the other hand, RMSE_B highlights worse edge preservation, because it is just evaluated only in subset B of pixels that have not been corrupted by noise. Ideally, these pixels should maintain their original values after noise removal. Indeed, RMSE_B is zero only if the original values of these pixels are preserved after filtering. The unwanted smoothing typically produced by denoising techniques is more relevant on object edges than on the uniform regions of the image. Thus, RMSE_B can effectively measure the edge preservation yielded by a filter.

VRMSE is also helpful when the perceived visual quality of two images is similar, as in the following example. At first, the image processed by the powerful 3×3 signal-dependent rank

Fig. 6. Result yielded by the 3×3 SD-ROM filter.

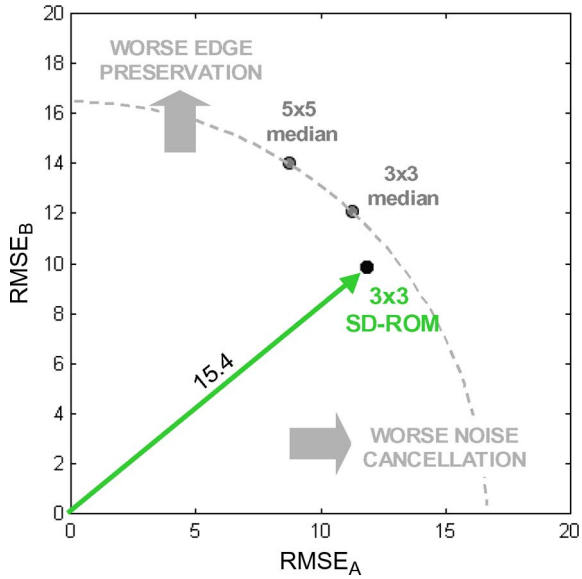


Fig. 7. Example of the graphical representation of the vector error for the 3×3 SD-ROM filter ("Boats" picture corrupted by "salt and pepper" impulse noise with a probability of 24.7 %).

ordered mean (SD-ROM) filter (Fig. 6) [30] does not seem to be very different from the result given by the simpler 3×3 median [Fig. 4(c)]. However, the different behaviors from the point of view of noise cancellation and edge preservation become more apparent when we plot the corresponding VRMSE values (Fig. 7). The smaller $|\text{VRMSE}|$ value yielded by the SD-ROM operator is clearly due to the fact that this filter is much better at preserving the image edges and slightly worse at removing noise pulses. A good filter showing a well-balanced combination of noise cancellation and edge preservation should give a VRMSE value located on the median of this graph with a low value.

B. Vector Approach for Images Corrupted by Gaussian Noise

The extension of the method to the IQA of data affected by other kinds of noise such as Gaussian (or uniform) noise

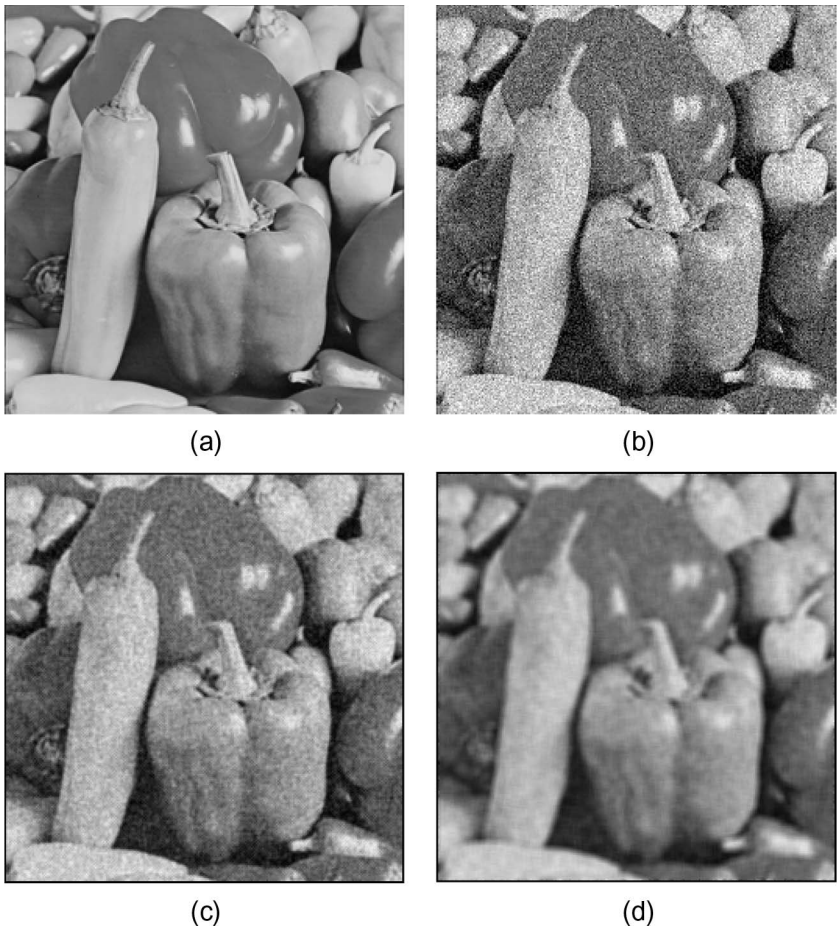


Fig. 8. (a) Original "Peppers" picture. (b) Image corrupted by Gaussian noise with standard deviation $\sigma = 26.5$. (c) Result yielded by the five-point arithmetic mean filter. (d) Result yielded by the 5×5 arithmetic mean filter.

requires a different approach. Indeed, when all data are noisy, we cannot deal with two different subsets of pixels formed by corrupted and uncorrupted data, respectively, as in the previous case. Thus, an appropriate tool should be devised to distinguish between noise cancellation and edge preservation. According to visual perception, a natural choice could be focusing on the uniform and nonuniform areas of a picture. In the following example, we shall exploit the well-known behavior of arithmetic mean filters to obtain two output pictures that have the same MSE but different image qualities. Let us consider the well-known 256×256 "Peppers" image [Fig. 8(a)]. The picture in Fig. 8(b) has been generated by adding zero-mean Gaussian noise with standard deviation $\sigma = 26.5$. The results given by two arithmetic mean filters with different window sizes are shown in Fig. 8(c) (five-point cross window) and Fig. 8(d) (5×5 window). Even if the processed images have the same MSE $= 162$, their different qualities are apparent. The large-window filter is more effective at reducing the noise, and this smoothing effect is well perceivable, particularly in the uniform (or slightly varying) regions of the image. However, the edges in the processed image are significantly blurred [Fig. 8(d)]. Not surprisingly, the small-window filter shows a complementary behavior. If we look at the object contours, we can notice that the blurring is less annoying. Of course, the filtered data remain very noisy. It should be observed that *uniform* or *slightly varying* regions clearly denote fuzzy concepts [31]. Thus, we shall resort to fuzzy membership functions to take into account how much a given pixel belongs to a uniform region or an object border. To numerically evaluate the aforementioned features, let us introduce the fuzzy VRMSE as follows:

$$\mathbf{VRMSE} = [\mathrm{RMSE}_{\tilde{A}}, \mathrm{RMSE}_{\tilde{B}}] \tag{6}$$

$$\mathrm{RMSE}_{\tilde{A}} = \sqrt{\frac{1}{N} \sum_{\mathbf{n}} \alpha(\mathbf{n}) \left[y(\mathbf{n}) - r(\mathbf{n})\right]^2} \tag{7}$$

$$\mathrm{RMSE}_{\tilde{B}} = \sqrt{\frac{1}{N} \sum_{\mathbf{n}} \beta(\mathbf{n}) \left[y(\mathbf{n}) - r(\mathbf{n})\right]^2} \tag{8}$$

where N is the total number of processed pixels in the region of interest S, and $\alpha(\mathbf{n})$ and $\beta(\mathbf{n})$ are the membership functions that describe the "uniform" and "not uniform" fuzzy sets, respectively [$0 \leq \alpha(\mathbf{n}) \leq 1$, $0 \leq \beta(\mathbf{n}) \leq 1$, $\beta(\mathbf{n}) = 1 - \alpha(\mathbf{n})$]. In our approach, $\beta(\mathbf{n})$ is typically obtained from a map of edge gradients of the reference image. The evaluation of the fuzzy VRMSE for the filtered pictures mentioned in the previous example is shown in Fig. 9 ($N = 252 \times 252$). In this experiment, we computed $\beta(\mathbf{n})$ by rescaling the output y_{sob} of the classical 3×3 Sobel operator to the interval [0, 1]: $\beta(\mathbf{n}) = y_{\mathrm{sob}}/(L-1)$, $(L = 256)$. The corresponding map

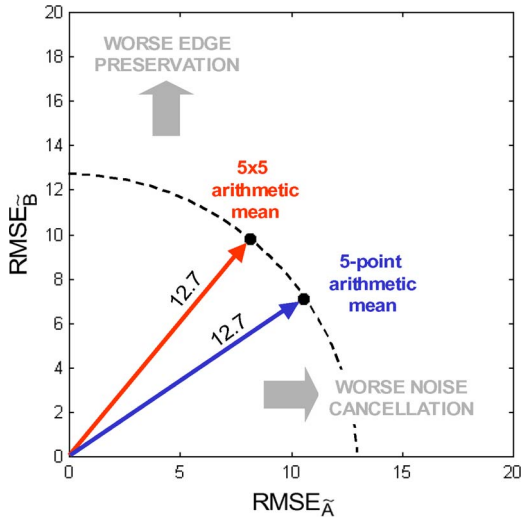
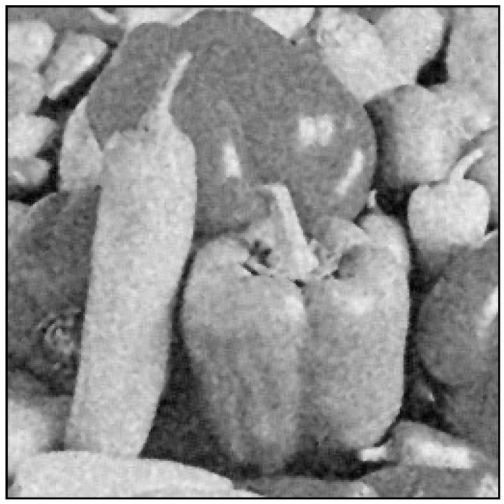
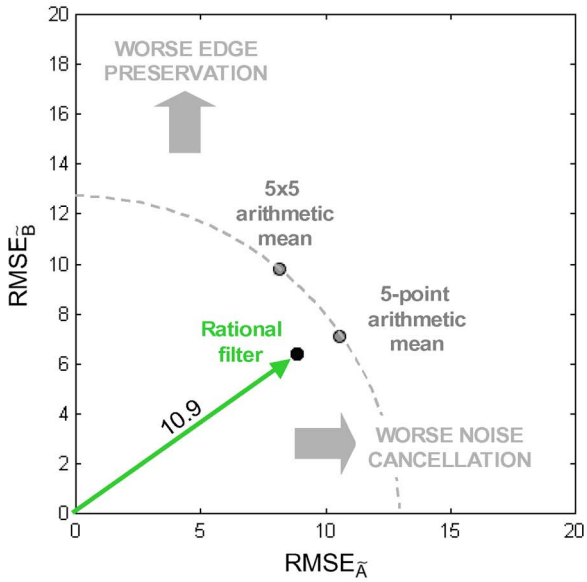


Fig. 9. Example of the graphical representation of the vector errors for the five-point and 5 × 5 arithmetic mean filters ("Peppers" picture corrupted by Gaussian noise with standard deviation $\sigma = 26.5$).

Fig. 10. Map of the edge gradients given by the Sobel operator.



(a)



(b)

Fig. 11. (a) Result yielded by the 3 × 3 rational filter. (b) Example of the graphical representation of the vector error for the 3 × 3 rational filter ("Peppers" picture corrupted by Gaussian noise with standard deviation $\sigma = 26.5$).

of edge gradients is shown in Fig. 10. Since the processed images have the same MSE, both vectors have the same length ($|\text{VRMSE}| = 12.7$). However, the different orientations of the vectors clearly show the different features of these linear filters, according to the preceding observations. Of course, nonlinear filters can give better results. As an example, the image processed by the effective 3 × 3 rational smoother [32] is shown in Fig. 11(a). The corresponding VRMSE is presented in Fig. 11(b) ($|\text{VRMSE}| = 10.9$), which shows the very good combination of noise cancellation and edge preservation given by this nonlinear filter. More examples dealing with test pictures corrupted by different amounts of Gaussian noise are shown in Fig. 12. The VRMSE evaluation offers a simple and effective way to assess the quality of the filtered images. If we look at the graphical representation of the vector error, we can immediately measure the noise still present in the data and the image degradation produced by the filtering action.

We have dealt with noise that is independent of spatial coordinates and uncorrelated with respect to the image itself. If some correlation between pixel values and the noise occurs, (6)–(8) still apply. Since the functions $\alpha(\mathbf{n})$ and $\beta(\mathbf{n})$ are obtained from a map of edge gradients of the reference (noise-free) picture, the $\text{RMSE}_{\tilde{A}}$ and $\text{RMSE}_{\tilde{B}}$ components yield, in any case, the filtering errors in the uniform and edge regions of the image, respectively.

Data denoising is a task of paramount importance for image-based instrumentation, because it can significantly reduce the uncertainties in subsequent measurement procedures, such as parameter estimation and object recognition. The development of criteria for the validation of the results is a key issue for the development of new filtering techniques. The use of two MSE components calculated for two subsets of the image pixels largely removes the limitations associated with the scalar MSE method used on the entire image. On the other hand, according to (5), the proposed method is backward compatible with the

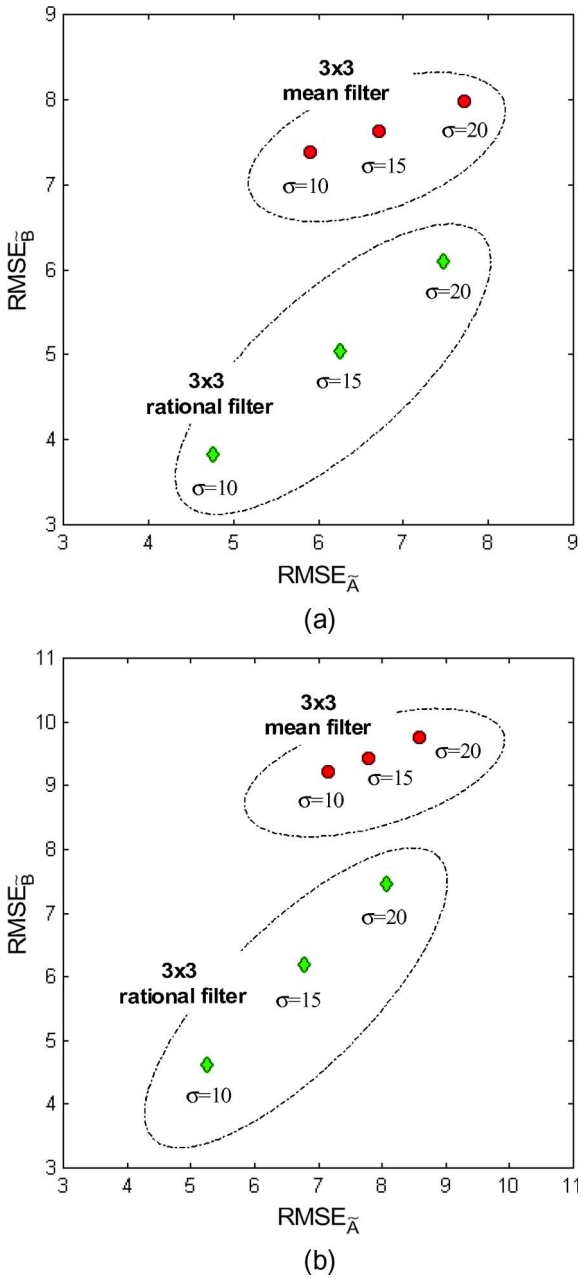


Fig. 12. Example of the graphical representation of the vector errors given by the 3×3 arithmetic mean and the 3×3 rational filter for the (a) "Lena" and (b) "Boats" pictures corrupted by three different amounts of Gaussian noise.

scalar MSE that is widely adopted in the literature. In principle, our approach is not limited to the adoption of the MSE since it could be implemented using any pixelwise metric. However, the MSE has the advantage of very simple implementation and low complexity. Furthermore, the proposed method lies in the framework of measurement science, where the concept of error as the difference between the result and the *true value* plays a very relevant role.

C. Extension of the Vector Approach to Color Images

We briefly discuss here the extension of the vector approach to color images. If we deal with Red Green Blue pictures, a straightforward way for such an extension is to perform the VRMSE evaluation for each image channel and to collect three vectors VRMSE_R, VRMSE_G, and VRMSE_B. However, the simplest way to extend our vector approach to color images is to choose a color model that guarantees compatibility with monochrome pictures, such as the YIQ color coordinate system [33]. In this model, the luminance is represented by the Y channel, whereas the I and Q components deal with the chrominance information only. Thus, we can define the VRMSE for color images $\text{VRMSE}_{\text{COLOR}}$ as follows:

$$\mathbf{VRMSE}_{\text{COLOR}} = [\text{RMSE}_Y, \text{RMSE}_C] \tag{9}$$

where the *luminance root mean squared error* RMSE_Y and the *chroma root mean squared error* RMSE_C are given by the following relationships:

$$\text{RMSE}_Y = \sqrt{\frac{1}{N}\sum_{\mathbf{n}} [y_1(\mathbf{n}) - r_1(\mathbf{n})]^2} \tag{10}$$

$$\text{RMSE}_C = \sqrt{\frac{1}{N}\sum_{\mathbf{n}} [y_2(\mathbf{n}) - r_2(\mathbf{n})]^2 + [y_3(\mathbf{n}) - r_3(\mathbf{n})]^2}. \tag{11}$$

In (10) and (11), $\{y_1, y_2, y_3\}$ and $\{r_1, r_2, r_3\}$ denote the Y, I, and Q components in the processed and reference images, respectively. RMSE_Y can be split into two components, as we did for monochrome images [see (6)–(8)], i.e.,

$$\mathbf{VRMSE}_Y = [\text{RMSE}_{\tilde{A}}, \text{RMSE}_{\tilde{B}}] \tag{12}$$

$$\text{RMSE}_{\tilde{A}} = \sqrt{\frac{1}{N}\sum_{\mathbf{n}} \alpha(\mathbf{n}) [y_1(\mathbf{n}) - r_1(\mathbf{n})]^2} \tag{13}$$

$$\text{RMSE}_{\tilde{B}} = \sqrt{\frac{1}{N}\sum_{\mathbf{n}} \beta(\mathbf{n}) [y_1(\mathbf{n}) - r_1(\mathbf{n})]^2}. \tag{14}$$

The validation of filters is not the only purpose of the proposed method. An application example dealing with the quality assessment of compressed color images is shown in Figs. 13–15. It is well known that blocking and ringing artifacts typically affect digital images after lossy compression, such as the JPEG technique. Blocking artifacts are represented by abrupt transitions of luminance across block boundaries. Ringing artifacts are perceivable as sharp oscillations or *ghost shadows* located along the object edges in the image. Thus, according to the preceding definitions, the proposed $\text{VRMSE}_{\text{COLOR}}$ is well suited for measuring the aforementioned effects. Fig. 13(a) shows a 256×256 slice of the 24-bit color image "Lena." The picture compressed at about 32 : 1 by means of the JPEG method is shown in Fig. 13(b). Ringing and blocking artifacts are apparent in the compressed image. Visual inspection shows that a better result can be given by the more powerful JPEG 2000 technique [Fig. 13(c)]. However, the $\text{VRMSE}_{\text{COLOR}}$ evaluation shown in Fig. 14 $(\text{VRMSE}_{\text{COLOR}})$ and Fig. 15 (VRMSE_Y) combines qualitative and quantitative

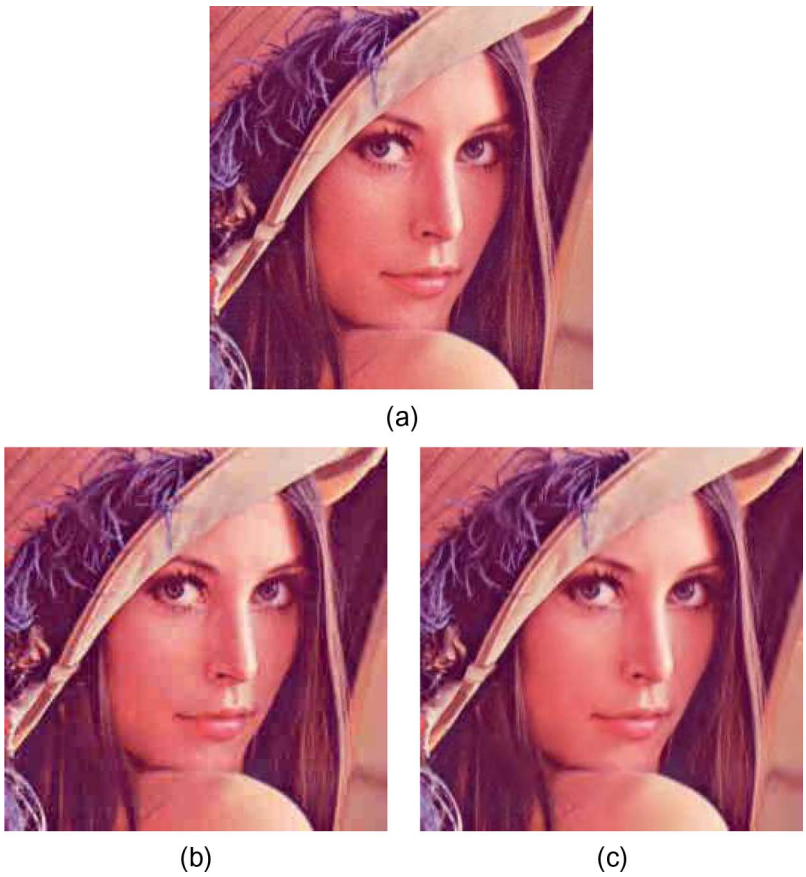


Fig. 13. (a) A 256 × 256 slice of the 24-bit color image "Lena" data compressed at about 32 : 1 by means of the (b) JPEG and (c) JPEG 2000 techniques.

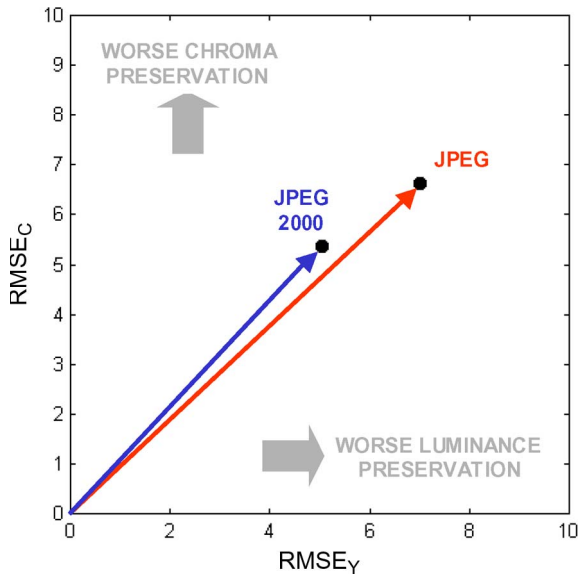


Fig. 14. Example of the graphical representation of the $\mathrm{VRMSE}_{\mathrm{COLOR}}$ vector error for the JPEG and JPEG 2000 techniques (24-bit color "Lena" picture compressed at about 32 : 1).

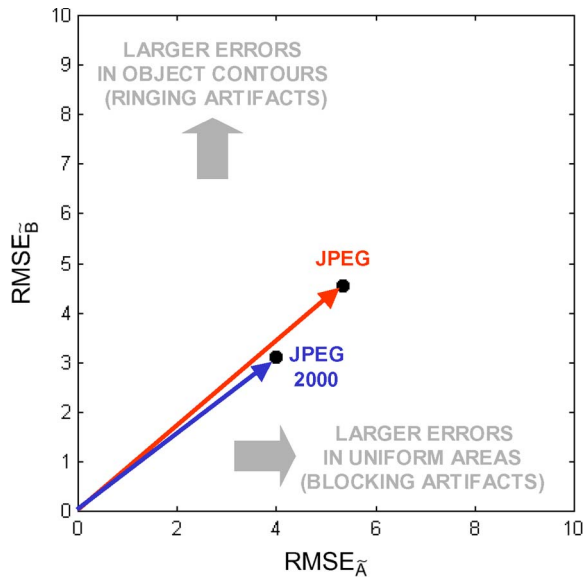


Fig. 15. Example of the graphical representation of the VRMSE_{Y} vector error for the JPEG and JPEG 2000 techniques (24-bit color "Lena" picture compressed at about 32 : 1).

information about the occurrence of different kinds of errors that visual inspection or classical MSE evaluation is unable to yield alone. Clearly, more attributes could be added to the image quality vector to take into account more features. A possible improvement, dealing with the evaluation of spatial granularity of the noise, will be a subject of future work.

VI. Conclusion

In this paper, a state of the art in the field of IQA has been traced, and a review of the main statistical performance evaluation methods has been provided. A comparison with the software quality measurement area has also been presented

because of the similarity that can be established with some issues in IQA. Furthermore, some metrological considerations on uncertainty evaluation and measurement modeling have been given. Finally, a vector approach for IQA has been proposed, introducing the VRMSE and fuzzy VRMSE methods, for the applications of images corrupted by impulsive and Gaussian noise, respectively. This approach may be used to characterize the behavior of a noise-suppression filter with respect to noise cancellation and detail preservation. A possible extension to color images has also been presented.

REFERENCES

[1] Z. Wang, A. C. Bovik, and S. Lu, "Why is image quality assessment so difficult?" in *Proc. IEEE ICASSP*, May 13–17, 2002, vol. 4, pp. IV-3313–IV-3316.

[2] F. Russo, "Automatic enhancement of noisy images using objective evaluation of image quality," *IEEE Trans. Instrum. Meas.*, vol. 54, no. 4, pp. 1600–1606, Aug. 2005.

[3] A. De Angelis, A. Moschitta, F. Russo, and P. Carbone, "Image quality assessment: An overview and some metrological considerations," in *Proc. IEEE Workshop AMUEM*, Sardagna, Italy, Jul. 16–18, 2007, pp. 47–52.

[4] I. E. G. Richardson, *H.264 and MPEG-4 Video Compression: Video Formats and Quality*. Hoboken, NJ: Wiley, Sep. 2003, ch. 2.

[5] *Methodology for the Subjective Assessment of the Quality of Television Pictures*, 2002. ITU-T Recommendation BT.500-11.

[6] Z. Wang, A. C. Bovik, and B. L. Evan, "Blind measurement of blocking artifacts in images," in *Proc. Int. Conf. Image Process.*, Sep. 10–13, 2000, vol. 3, pp. 981–984.

[7] P. Gastaldo and R. Zunino, "No-reference quality assessment of JPEG images using CBP neural networks," in *Proc. ISCAS*, May 23–26, 2004, vol. 5, pp. V-772–V-775.

[8] P. Marziliano, F. Dufaux, S. Winkler, and T. Ebrahimi, "Perceptual blur and ringing metrics: Application to JPEG2000," *Signal Process., Image Commun.*, vol. 19, no. 2, pp. 163–172, Feb. 2004.

[9] Z. Wang, G. Wu, H. R. Sheikh, E. P. Simoncelli, E.-H. Yang, and A. C. Bovik, "Quality aware images," *IEEE Trans. Image Process.*, vol. 15, no. 6, pp. 1680–1689, Jun. 2006.

[10] I. Avcibas, B. Sankur, and K. Sayood, "Statistical evaluation of image quality measures," *J. Electron. Imaging*, vol. 11, no. 2, pp. 206–223, Apr. 2002.

[11] D. Van der Weken, M. Nachtegael, and E. E. Kerre, "Using similarity measures and homogeneity for the comparison of images," *Image Vis. Comput.*, vol. 22, no. 9, pp. 695–702, Aug. 2004.

[12] S. Daly, "The visible differences predictor: An algorithm for the assessment of image fidelity," in *Digital Images and Human Vision*, A. B. Watson, Ed. Cambridge, MA: MIT Press, 1993, pp. 179–206.

[13] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[14] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 403–444, Feb. 2006.

[15] H. R. Sheikh, A. C. Bovik, and G. de Veciana, "An information fidelity criterion for image quality assessment using natural scene statistics," *IEEE Trans. Image Process.*, vol. 14, no. 12, pp. 2117–2128, Dec. 2005.

[16] M. Miyahara, K. Kotani, and V. Ralph Algazi, "Objective Picture Quality Scale (PQS) for image coding," *IEEE Trans. Commun.*, vol. 46, no. 9, pp. 1215–1226, Sep. 1998.

[17] A. B. Watson, "DCTune: A technique for visual optimization of DCT quantization matrices for individual images," in *Soc. Inf. Display Dig. Tech. Papers XXIV*, 1993, pp. 946–949.

[18] Video Quality Experts Group, *Final Report on the Validation of Objective Models of Video Quality Assessment*, Mar. 2000.

[19] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Trans. Image Process.*, vol. 15, no. 11, pp. 3441–3452, Nov. 2006.

[20] V. Petrovic, "Subjective tests for image fusion evaluation and objective metric validation," *Inf. Fusion*, vol. 8, no. 2, pp. 208–216, Apr. 2007.

[21] ISO/IEC, *Software Engineering—Product Quality—Part 1—Quality Model*, 2001, Genève, Switzerland: Int. Org. Stand., 9126-1:2001.

[22] P. Carbone, L. Buglione, L. Mari, and D. Petri, "Metrology and software measurement: A comparison of some basic characteristics," in *Proc. 23rd IEEE IMTC*, 2006, pp. 1082–1086.

[23] S. H. Kan, *Metrics and Models in Software Quality Engineering*. Reading, MA: Addison-Wesley, 2003.

[24] *Guide to the Expression of Uncertainty in Measurement*, ISO, Genève, Switzerland, 1995.

[25] H. R. Sheikh, Z. Wang, L. Cormack, and A. C. Bovik, *LIVE Image Quality Assessment Database Release 2*. [Online]. Available: http://live.ece.utexas.edu/research/quality

[26] M. G. Cox and P. M. Harris, *NPL report—Software Support for Metrology. Uncertainty Evaluation*, Mar. 2004. CMSC 40/04.

[27] R. Aldridge, J. Davidoff, M. Ghanbari, D. Hands, and D. Pearson, "Recency effect in the subjective assessment of digitally-coded television pictures," in *Proc. 5th Int. Conf. IEEE Image Process. Appl.*, pp. 336–339.

[28] D. S. Hands, M. D. Brotherton, A. Bourret, and D. Bayart, "Subjective quality assessment for objective quality model development," *Electron. Lett.*, vol. 41, no. 7, pp. 408–409, Mar. 31, 2005.

[29] C. Glasman, V. Andronov, A. Bukina, and O. Vasilyev, "Subjective assessment of compression systems by trained and untrained observers," in *Proc. Int. Broadcast. Conv.*, Sep. 12–16, 1997, pp. 476–481.

[30] E. Abreu, "Signal-dependent rank-ordered mean (SD-ROM) filter," in *Nonlinear Image Processing*, S. K. Mitra and G. L. Sicuranza, Eds. San Diego, CA: Academic, 2001, pp. 111–134.

[31] F. Russo, "Fuzzy model fundamentals," in *Wiley Encyclopedia of Electrical and Electronics Engineering*, J. G. Webster, Ed. Hoboken, NJ: Wiley, 1999 (updated 2007).

[32] G. Ramponi, "Polynomial and rational operators for image processing and analysis," in *Nonlinear Image Processing*, S. K. Mitra and G. L. Sicuranza, Eds. San Diego, CA: Academic, 2001, pp. 203–224.

[33] I. Pitas, *Digital Image Processing Algorithms and Applications*. Hoboken, NJ: Wiley, 2000.

Alessio De Angelis (S'07) was born in Perugia, Italy, in 1979. He received the Laurea degree in electronic engineering, with a thesis work on the digital signal processing and digital video compression areas, from the University of Perugia in 2004. He is currently working toward the Ph.D. degree in information engineering with the Department of Electronic and Information Engineering, University of Perugia.

His current research interests include indoor geolocation, short-pulse ultrawideband systems, and image quality assessment.

Antonio Moschitta (S'94–M'02) was born in Foligno, Italy, in 1972. He received the Laurea degree in electronic engineering, with a thesis focused on the performance assessment of digital terrestrial television transmitters, and the Ph.D. degree in electronic engineering, with his final dissertation focused on the effects of quantization noise on the performances of digital communication systems, from the University of Perugia, Perugia, Italy, in 1998 and 2002, respectively.

He is currently an Assistant Professor with the Department of Electronic and Information Engineering, University of Perugia. His research interests include modern digital communication systems, analog-to-digital and digital-to-analog conversion, Sigma–Delta converters, and parametric estimation.

Fabrizio Russo (M'88–SM'06) received the Dr. Ing. degree (*summa cum laude*) in electronic engineering from the University of Trieste, Trieste, Italy, in 1981.

In 1984, he joined the Dipartimento di Elettrotecnica Elettronica Informatica (D.E.E.I.), University of Trieste, where he is currently an Associate Professor of electrical and electronic measurements. He is the author of more than 100 papers in international journals, textbooks, and conference proceedings. His research interests include nonlinear signal processing and the application of computational intelligence to instrumentation including fuzzy and neurofuzzy techniques for image filtering, noise estimation, and edge detection.

Prof. Russo was one of the organizers of the 2004 and 2005 IEEE International Workshops on Imaging Systems and Techniques. He served as a Technical Program Co-Chairman of the 2006 IEEE Instrumentation and Measurement Technology Conference and as a Co-Guest Editor for the IEEE Instrumentation and Measurement Technology Conference 2006 Special Issue of the IEEE TRANSACTIONS ON INSTRUMENTATION AND MEASUREMENT, which was published in August 2007.

Paolo Carbone (S'91–A'95) received the Laurea and the Ph.D. degrees from the University of Padova, Padova, Italy, in 1990 and 1994, respectively.

From 1994 to 1997, he was a Researcher with the Third University of Rome, Rome, Italy. Since 2002, he has been a Full Professor with the Department of Electronic and Information Engineering, University of Perugia, Perugia, Italy, where he teaches courses in instrumentation and measurement and in reliability and quality engineering. He has been involved in various research projects sponsored by the Italian Government. He is the author/coauthor of several papers in international journals and conference proceedings. His research objective is to develop knowledge, models, and systems for the advancement of instrumentation and measurement technology. His research involves the development of original techniques for signal acquisition analysis and interpretation, with emphasis on the performance improvement of electronic instrumentation and data-acquisition systems using state-of-the-art technologies.

Prof. Carbone was the Chairman of the 8th International Workshop on ADC Modelling and Testing (Perugia, Sep. 8–10, 2003). He served as an Associate Editor for the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS—PART II from 2000 to 2002 and as an Associate Editor for the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS—PART I from 2005 to 2007.