Int J Digit Libr (2014) 15:1–25
DOI 10.1007/s00799-014-0124-0

How to assess image quality within a workflow chain: an overview

Gianluigi Ciocca · Silvia Corchs · Francesca Gasparini · Raimondo Schettini

Received: 7 February 2013 / Revised: 1 August 2014 / Accepted: 4 August 2014 / Published online: 15 August 2014

Abstract Image quality assessment (IQA) is a multidimensional research problem and an active and evolving research area. This paper aims to provide an overview of the state of the art of the IQA methods, putting in evidence their applicability and limitations in different application domains. We outline the relationship between the image workflow chain and the IQA approaches reviewing the literature on IQA methods, classifying and summarizing the available metrics. We present general guidelines for three workflow chains in which IQA policies are required. The three workflow chains refer to: high-quality image archives, biometric system and consumer collections of personal photos. Finally, we illustrate a real case study referring to a printing workflow chain, where we suggest and actually evaluate the performance of a set of specific IQA methods.

Keywords Image quality assessment · Image quality metrics · Image production workflow chain · Printing workflow chain

1 Introduction

Images play a more and more important role in sharing, expressing, mining and exchanging information in our daily lives. Now, we can all easily capture and share images anywhere and any time. Since digital images are subject to a wide variety of distortions during acquisition, processing, compression, storage, transmission and reproduction, it becomes necessary to have tools that make it possible to assess the image quality (IQ) during the whole production chain. This can be done interactively by subjective human rating or automatically by objective methods. Different definitions of quality are found in the literature [25,48,55,103,123]. According to the International Imaging Industry Association [44], IQ is the perceptually weighted combination of all visually significant attributes of an image when considered in its marketplace or application. This is also stressed by the Technical Advisory Service for Images: "The quality of an image can only be considered in terms of the proposed use. An image that is perfect for one use may well be inappropriate for another". While many different methods for IQA have been proposed in the literature, there is a lack of analysis and discussion about when and which of these methods can be used within a specific image workflow chain. Let us, for example, focus on an image workflow chain like the one that starts with a scene to be captured by a imaging device and published on the web. Which of the available methods should be applied and at what precise point of the chain, to get an estimation of the quality of the final output image?

Many IQ metrics can be found in the literature, and in this overview, we listed more than 50 of them. So, what metric should be used for a given task? It is not easy to give a definitive answer to such question since it depends on different factors: the semantic content of the image, the application task, and the particularly imaging chain applied. This paper is aimed at users who want to have a broad overview of the available metrics, and insights into their applications in different domains. To this end, we provide a compendium

G. Ciocca (✉) · S. Corchs · F. Gasparini · R. Schettini
DISCo, University of Milano Bicocca, Milan, Italy
e-mail: ciocca@disco.unimib.it

S. Corchs
e-mail: corchs@disco.unimib.it

F. Gasparini
e-mail: gasparini@disco.unimib.it

R. Schettini
e-mail: schettini@disco.unimib.it

of the state of the art of the different IQA methods. We classify and summarize the different available metrics outlining the relationship between the image workflow chain and the image quality assessment (IQA) approaches. We show how and when these different kinds of metrics can be applied within a generic image workflow chain and within specific application scenarios. Finally, several open issues currently being addressed by the IQA community are presented and discussed. We hope that, at the end, users will have: a more clear view of the different faces of IQA, a comprehension of the characteristics of the available metrics, and the tools to make the proper choices for their tasks.

2 Modeling image quality

Some attempts have been made in the last decade to develop a general, broadly applicable, IQ model that regards images not only as signals but also as carriers of visual information, which encode important and useful information about the geometry of the scene and the properties of the objects located within this scene [53,108,132].

Different properties contribute to define image quality and different models have been proposed in the literature. The fidelity-usefulness-naturalness (FUN) IQ model [86] assumes the existence of these three major dimensions:

- Fidelity is the degree of apparent match of the image with the original. Ideally, an image having the maximum degree of fidelity should give the same impression to the viewer as the original. As an example, a painting catalogue requires high fidelity of the images with respect to the originals. Genuineness and faithfulness are sometimes used as synonyms of fidelity [44]. Dozens of books and thousands of papers have been written about image fidelity and image reproduction, e.g. [94].
- Usefulness is the degree of apparent suitability of the image with respect to a specific task. In many application domains, such as surveillance, automotive, medical or astronomical imaging, image processing procedures can be applied to increase the image usefulness [41].
  The enhancement processing steps have an obvious impact on fidelity.
- Naturalness is the degree of apparent match of the image with the viewer's internal references. This attribute plays a fundamental role when we have to evaluate the quality of an image without having access to the corresponding original. Examples of images requiring a high degree of naturalness are photos downloaded from the web, or seen in journals. Naturalness also plays a fundamental role when the image to be evaluated does not exist in reality, such as in virtual reality domains.

Recently, Moorthy et al. [71] suggested extending the dimensions of image quality by considering also its *Visual Aesthetic* and *Content*. We may refer to their model as the QAC model (quality, aesthetic, content).

- Visual aesthetics refers to the perceived beauty of an image. Aesthetics is intrinsically subjective; different users may consider an image to be aesthetically appealing for different motives based on their backgrounds and expectations. Notwithstanding the subjective nature of this dimension, several works tackle the problem to estimate the aesthetic of a photo by developing automatic computational procedures. These procedures exploit visual properties and compositional rules trying to predict aesthetic scores with high correlation with human perception [7,30,75].
- Semantic content has an important impact on the evaluation of the quality of an image and thus it cannot be discounted during assessment. The application in hand and users' previous experiences influence the judgment of a good or bad image content. An image can be considered good if all the relevant (for the user) content is clearly visible or if the image conveys the expected information. On the contrary, an image can be considered of poor quality if it depicts offensive or disgusting (for the user) content.

3 Image quality assessment approaches

Different criteria can be used to classify the IQA approaches. At the top level, we may divide the methods into two major groups:

- Approaches that take into account the quality of the image itself. We may call these approaches, *"direct"* approaches (Engeldrum [32] calls them "beauty contest models") and are used when the images themselves must be compared with other ones, either for competition or reference (see section 3.1);
- Approaches that consider the quality of the images with respect to the performances reached by the application that uses them. We may call these approaches, *"indirect"* approaches (in [32] are termed "detection/recognition models") and are also used to evaluate the whole system, device or algorithm that process the image (see Sect. 3.2).

3.1 Direct image quality approaches

Direct image quality approaches can be categorized in subjective versus objective methods. Subjective methods are based on psychological experiments involving human observers. Different techniques can be used, and the

methods indicating how to perform subjective quality assessment are described in some standards, such as ITU-R BT.500-13 [52,112]. Objective methods compute suitable metrics directly from the digital image without human observers. These objective methods can be further classified according to many different criteria depending on the available data and the type of assessment output.

3.1.1 Subjective approaches

The involvement of real people who view the images to assess their quality requires that all the factors that influence perception are taken into account to discount possible biases, To this end, strict protocols have to be adopted. In the ITU standards [52], different subjective test methodologies are described. Regardless of the choice of the test methodology used, the way in which responses of the tests are analysed depends upon the judgment (e.g. detection) and the information sought.

Test methods can be grouped into two main categories: methods that use explicit references, and methods that do not use any explicit reference. Single Stimulus (SS) methods belong to the first category, while Stimulus Comparison (SC) methods belong to the second one. In SS methods, a single image or sequence of images is presented and the assessor provides a quality score of the presentation, while in SC methods, two images or set of images are displayed, and the viewer provides a rating of the relation among the images.

For both SC and SS methods, there are different variants and the main difference is in the scale that the assessors use to evaluate the presentations. For example, in adjectival categorical judgments, observers assign an image or image sequence to one of a set of categories that, typically, are defined in semantic terms. The categories may reflect judgments about the existence of a perceptible difference (e.g. "SAME", "DIFFERENT") or the existence and direction of perceptible differences (e.g. "LESS", "SAME", "MORE"). Categorical scales that assess image quality and image impairment have been used most often, and in [52] readers can find some suggested scales to be used in the evaluation process. For each attribute/artifact, this method yields a distribution of judgments across scale categories.

In non-categorical judgments, observers assign a numerical value to each image or image sequence shown. These methods can have two kinds of scales: continuous or discrete. In continuous scaling, a variant of the categorical method, the assessor assigns each image or image sequence to a point on a line drawn between two semantic labels. The distance from an end of the scale is taken as the index for each presentation. In discrete scaling, the assessor assigns each image or image sequence a number that reflects its judged level on a specified dimension (e.g. image sharpness). The range of the numbers used may be restricted (e.g. 0–100) or not. Sometimes, the number assigned describes the judged level in absolute terms without direct reference to the level of any other image or image sequence as in some forms of magnitude estimation. In other cases, the number describes the judged level relative to that of a reference [56].

An important variant of the Stimulus Comparison is the Pairwise Comparison (PC) which is based on the law of comparative judgment studied by Thurstone [106]. In the PC method, the images are organized in sequences, each of which usually contains different versions of the same image. The images in each sequence are presented in pairs in two locations (for example one on the left and one on the right of the display) in all the possible combinations. Thus, each image is displayed twice in both locations. After each pair is presented, a judgment is made on which element in the pair is preferred based on some attributes. In PC, the tester does not impose any scale for the assessment. The selection of one image over the other is an exclusive Boolean choice.

The obtained scores can then be used as such, normalized using the mean and standard deviation to obtain Z scores, or Thurstone scaling [106] can be used to create an interval scale, so that the scale represents equal perceptual distances. Finally, the quality ratings from the evaluators are averaged to obtain the Mean Opinion Score (MOS) or the Difference Mean Opinion Score (DMOS). The latter is the difference between the MOS scoring of the test image and the MOS scoring of the corresponding reference image.

Notwithstanding the effectiveness of subjective approaches, their efficiency is very low.

3.1.2 Objective approaches

Depending on the availability of the original image, the direct image quality approaches can be categorized into three groups: Full Reference (FR), No Reference (NR) and Reduced Reference (RR) [9]. Taking into account the philosophy followed when constructing their algorithms, the methods can be also classified as bottom-up or top-down. If we consider the application scope, they can be general purpose or context dependent. In the following, we summarize the well-known metrics belonging to the different FR, NR and RR categories.

3.1.3 Full-reference metrics

Full-reference (FR) metrics (see Fig. 1 straight line paths) perform a direct comparison between the image under test and a reference or "original" in a properly defined image space. These methods are the first choice for subjective IQA. Having access to an original is a requirement of the usability of such metrics. Among the quality dimensions previously introduced, only image fidelity is assessed. The sim-

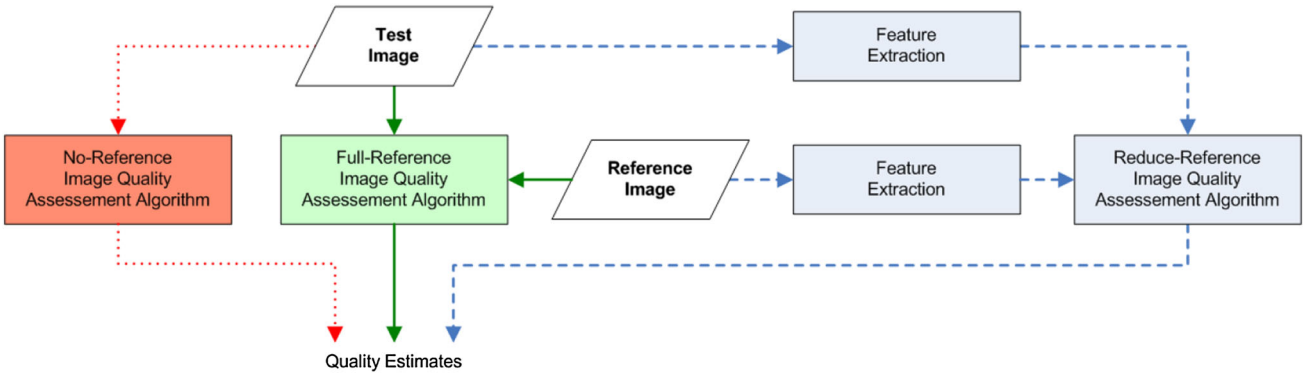

Fig. 1 Objective image quality assessment approaches

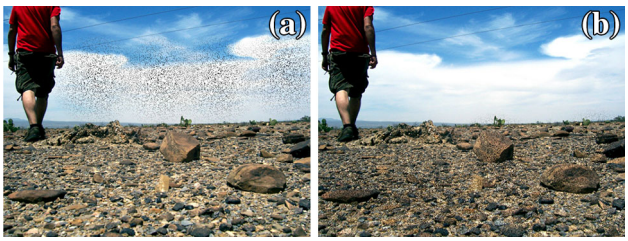

Fig. 2 Example of how the perceptual quality is influenced by the visibility of the distortion. Gaussian noise is applied to the *top* (**a**) and *bottom* (**b**) regions of the image. The image in **b** is perceived as having higher quality than the image in **a**

plest FR metric is the Mean Square Error (MSE) or Peak to Signal Noise Ratio (PSNR). Even if they are the most used, in general they do not correlate with subjective assessments [40,113].

Consider, for example, Fig. 2. The same amount of Gaussian noise is applied to the image, first on the sky/clouds region and then on the sand/rocks one (Fig. 2a, b, respectively). The perceived image quality is influenced by the distortion visibility: when the distortion is applied to the sand/rock region, it is less noticeable as the noise is masked by the variations in the texture of the region. When the distortion is applied to almost uniformly regions, as in the case of the sky/clouds region, it stands out prominently. This effect is called *Noise Masking* or *Texture Masking* [104,122] and it is fundamental to take it into account when designing image quality metrics.

Error sensitivity frameworks follow a strategy of modifying MSE-like measures so that errors are penalized in accordance with their visibility. The evaluation of the visibility is accomplished by modeling some aspects of the human visual system (HVS) like Channel Decomposition, Contrast Sensitivity and Point Spread functions among others [29,64,89,105]. All these techniques are bottom-up like approaches.

Top-down approaches take into account, for example, the image structure in defining the IQ since they assume that finding the structure is the goal for the cognitive process. The structural information in an image is defined as those attributes that represent the structure of objects in the scene, independently of the average luminance and contrast. The image quality is measured as a function of the amount of distortion that influences the image structure. To cope with the fact that the human visual system processes a scene at different level of details, some structural-based metrics have been also extended to process an image with a multi-scale approach. Quality at different scales contributes differently to the overall quality. Other approaches consider the characteristics of the natural images. They use natural scene statistics to quantify the loss of information due to the distortions present in the image.

A brief summary of FR metrics is presented in Table 1. The performance of each metric in terms of correlation with a ground truth, if available, is reported in the last column of the table (see Sect. 3.3).

3.1.4 No-reference (NR) metrics

No-reference (NR) metrics (see Fig. 1 dotted line paths) are also called blind metrics and assume that IQ can be determined without a direct comparison between the original and the processed images. NR metrics can be used whenever the original image is unavailable. While the FR metrics estimate the image quality in an holistic way, NR metrics are often targeted to estimate the presence of a specific image defect introduced by the imaging device, some image processing procedures or by the transmission channel. This implies that some information about the application requirements and users' preferences are needed to contextualize the quality measures. The overall image quality can be also evaluated using an ensemble of different NR metrics. Different types of defects can be considered [43]: blurriness, the attenuation of the high spatial frequencies; blocking, discontinuities generated by block-based compression algorithms such as JPEG; graininess, random fluctuation of pixel values due to

Table 1 Full Reference methods

FR method	Brief description	Performance
MSE, PSNR	Measures the fidelity to the original and does not take into account HVS characteristics. It is the simplest and oldest measure. No parameters are needed	–
Error sensitivity frameworks Daly [29], Lubin [64], Safranek and Johnston [89], Teo and Heeger [105], Waston [121] (1989–1993)	Measure the fidelity to the original. These are bottom-up approaches that simulate functional properties of the HVS. Consist essentially in four modules: preprocessing (alignment, luminance transformation, and color transformation), channel decomposition (different choices are identity, wavelet, discrete cosine and gabor transform), error weighting and error summation (Minkowski error pooling). Different parameters have to be estimated	–
Spatial-CIELAB Zhang and Wandell [135]	Measures color differences and is an extension of the CIELAB color metric. The image data are transformed into an opponent color space, followed by a CSF spatial filtering. An error map is evaluated. Different parameters have to be estimated	–
MultiScale Structural Similarity Index (MS-SSIM) Wang et al. [118]	An extension of the SSIM. Supplies more flexibility than previous single-scale methods in incorporating the variations of viewing conditions and image details	LIVE JPEG/JPEG2000 database PCC = 0.969 RMSE = 4.91 OR = 0.016
Structure Similarity Index (SSIM) Wang et al. [119]	Measures the fidelity to the original. The HVS is adapted to extract structural information from natural visual scenes. Models image degradation as structural distortions instead of errors. The SSIM index is obtained as the product of three comparison components: luminance, contrast and correlation. Different parameters have to be estimated. Is a top-down approach [9, chap. 3]	LIVE database PCC = 0.967 RMSE = 5.06 OR = 0.041 SROCC = 0.963
Visual Information Fidelity Index (VIF) Sheikh and Bovik [96]	Measures the information shared between the two images. The construction of the VIF Index relies on the modeling of the statistical image source, the image distortion channel, and the human visual distortion channel. Different parameters have to be estimated	LIVE database PCC = 0.949 RMSE = 5.083 OR = 0.013 SROCC = 0.949
Gradient structural similarity (G-SSIM) Chen et al. [16]	Image quality assessment is addressed by following the HVS's characteristic that human eye is very sensitive to the edge and contour information of an image, and the edge and contour information is the most important structural information of images	LIVE database PCC = 0.917 RMSE = 6.284 OR = 0.055
Most apparent distortion metric Larson and Chandler [59]	Combines two different strategies. For high-quality images, local luminance and contrast masking are used to estimate detection-based perceived distortion. On the other hand, changes in the local statistics of spatial-frequency components are used to estimate appearance-based perceived distortion in low-quality images	TID, LIVE, Toyama, and CSIQ databases PCC = 0.8306 (TID), 0.9683 (LIVE), 0.8951 (Toyama), 0.9502 (CSIQ) SROCC = 08340 (TID), 0.9675 (LIVE), 0.8908 (Toyama), 0.9466 (CSIQ)
Divisive normalization metric Laparra et al. [58]	Measures the closeness to the original. The metric is based on divisive normalization models [105] within discrete cosine transform and wavelet domains	–
Discrete orthogonal moments Wee et al. [124]	Measures the Moment Correlation Index. Up to fourth order moments are computed on non-overlapping blocks for both the test and reference images. Correlation indexes are computed on each pair of block moments, and a single-quality score is obtained by averaging all the correlation indexes. Two metrics are proposed: Q1 and Q2	LIVE, A57, IVC and MICT databases Q1: PCC = 0.608–0.937 SROCC = 0.606–0.947 Q2: PCC = 0.680–0.934 SROCC = 0.726–0.938
4-component gradient structural similarity (4-G-SSIM) Li and Bovik [60]	A four-component image model is used to classify image local regions according to edge and smoothness properties. SSIM scores are weighted by region type, leading to modified versions of the original G-SSIM index	LIVE database SROCC = 0.9594, PCC = 0.9491, RMSE = 5.0016

Table 1 continued

FR method	Brief description	Performance
Information content-weighted structural similarity measure (IW-SSIM) Wang and Li [114]	Information content weighting method built upon advanced statistical image models and combined with multiscale SSIM. The rationale is that when viewing natural images, the optimal perceptual weights for IQ pooling should be proportional to local information content, which can be estimated in units of bit using advanced statistical models of natural images	LIVE, TID2008, IVC, Toyama, A57, and CSIQ databases PCC = 0.9126 (average) SROCC = 09063 (average)
Feature Similarity Index (FSIM) Zhang et al. [134]	An image quality map is obtained by combining phase congruency (dimensionless measure of the significance of a local structure) and gradient magnitude (contrast information). After obtaining the local quality map, phase congruency is used as a weighting function to derive a single quality score	TID2008, CSIQ, LIVE, IVC, A57, and MICT databases SROCC = 0.8805 (TID2008)–0.9634 (LIVE) PCC = 0.8738 (TID2008)–0.9597 (LIVE)
Perceptual image quality assessment (PIQA) Fei et al. [33]	A luminance comparison measure, a structure comparison measure, and a contrast comparison measure are pooled together taking into account the contrast masking and neighborhood masking effects of the HVS perceptual process	LIVE, IVC and MICT databases PCC = 0.9024–0.9651 SROCC = 0.8890–0.9612
Machine learning approach Charrier and Lebrun [11]	It is based on a learned classification process to respect human observers. Support vector machine is applied for both classification and regression tasks. The feature vector contains visual attributes describing the images content	LIVE and TID2008 databases SROCC = 0.96 (LIVE), 0.90 (TID2008)

Fig. 3 Examples of image defects detected by no-reference metrics

the device sensor; contrast, the difference in the brightness that makes an object in an image distinguishable from other objects and the background; colorfulness, the perceived difference between a color and gray. Figure 3 shows some of these defects. Please note that the defects have been accentuated for ease of readability.

Blind methods can be classified as application dependent since they are defined to handle with one or few specific defect types. Some of the blind methods are carried out in the frequency domain (like [20] for example) and make use of the common statistical characteristics of the power spectra of natural images [109] to define the corresponding quality metrics. A variety of statistical properties of natural images (intensity, color, spatial correlation and higher order statistics) and their relationship to visual processing has been extensively studied by [98]. A brief summary of different NR methods is presented in Table 2. When possible we reported the overall performance score in the last columns. If this score is not available, we report some performance scores of the most common defects either as a single value or as a range of values.

3.1.5 Reduced-reference (RR) metrics

Reduced-reference (RR) metrics (see Fig. 1 dashed line paths) lie between FR and NR metrics. They are designed to predict perceptual IQ with only partial information about the reference image. In their general forms, these methods extract a number of features from both the reference and the image under test, and image comparison is based only on the correspondence of these features. Therefore, only image fidelity can be assessed. RR metrics are commonly used to track the

Table 2 No Reference methods

NR method	Artifacts	Brief description	Performance
Peli [78]	Contrast	Assigns a contrast value to every point in the image as a function of the spatial frequency band. The contrast is defined as the ratio of the bandpass-filtered image at that frequency to the low-pass image filtered to an octave below the same frequency (local luminance mean)	–
Immerkaer [47]	Noise	Estimates variance of the normally distributed noise	–
Wu and Yuen [126]	Blockiness	Generalized block-edge impairment (GBIM) metric, measures the blockiness separately in horizontal and vertical direction, after which the two directions are combined into a single-quality value. The GBIM assumes that the artifacts occur on a grid of blocks of pixels, which is common for most compression standards	LIVE database PCC = 0.9562, SROCC = 0.9522
Rank et al. [82]	Noise	Assumes Gaussian distributed noise. Estimates the noise variance. First, the noisy image is filtered by a horizontal and a vertical difference operator, then the histogram of local signal variances is computed. The mean square value of the histogram gives a noise estimation value	–
Vlachos [111]	Blockiness	Designed in the frequency domain. The blockiness measure is defined as the ratio between intra- and inter-block similarity	–
Wang et al. [116]	Blockiness	Defined in the frequency domain. They model the blocky image as a non-blocky image interfered with a pure blocky signal. The task of the blocking effect measurement algorithm is to detect and evaluate the power of the blocky signal. Luminance and texture masking effects are incorporated	–
Bovik and Liu [8]	Blockiness	Discrete cosine transform-domain algorithm. Blocking artifact modelled as a 2-D step function. Luminance and texture masking taken into account	–
Wang et al. [117]	Blockiness	Feature extraction method in the spatial domain. Measures differences across block boundaries and zero-crossings. Non linear regression is applied where the parameters are estimated from subjective tests	LIVE database RMSE = 7.76 PCC = 0.970 SROCC = 0.960
Marziliano et al. [66]	Blur	Defined in the spatial domain. An edge detector is applied. For pixels corresponding to an edge location, the start and end positions of the edge are defined as the local extrema locations closest to the edge. The edge width is measured and identified as the local blur measure. Global blur obtained by averaging the local blur values over all edge locations	105 images from LIVE and other sources PCC = 0.85–0.96 SROCC = 0.87–0.96
Corner et al. [24]	Noise	Laplacian and gradient data masks are used to estimate the additive and multiplicative noise standard deviations in an image. The histogram median value supplied the most accurate final noise estimations	–
Hasler and Süsstrunk [43]	Colorfulness	Study of the distribution of the image pixels in the CIELab color space, assuming that the colorfulness can be represented by a linear combination of a subset of different quantities (standard deviation and mean of saturation and/or chroma). Parameters are found by maximizing the correlation between experimental data and the metric	84 images PCC = 0.871–0.942
Ong et al. [76]	Blur	The average edge spread in the image is measured by the average extent of the slope spread of an edge in both the gradients' direction and also the direction opposing the gradients' direction	624 images RMSE = 0.1774
Pan et al. [77]	Blockiness	Measures horizontal and vertical inter-block difference. Takes into account the blocking artifacts for high bit rate images and the flatness for the very low bit rate images	LIVE database PCC = −0.930 SROCC = 0.932
Wang et al. [120]	Blur	Defined in the frequency domain. Blur is interpreted as a disruption of the local phase. The measure of phase coherence is based on coarse-to-fine phase prediction. The computations bear some resemblance to the behaviors of neurons in the primary visual cortex of mammals	–

Table 2 continued

NR method	Artifacts	Brief description	Performance
Winkler and Süsstrunk [125]	Noise	Investigates the visibility of noise itself as a target and uses natural images as the masker. Targets are Gaussian white noise and band-pass filtered noise of varying energy. Psychophysical experiments are conducted to determine the detection threshold of these noise targets on many different types of image content (noise visibility)	30 images PCC = 0.95
Muijs and Kirenko [72]	Blockiness	The key algorithm is based on the principle that block discontinuities can be characterized as edges that stand out from the spatial activity in their vicinity. The visibility of a block edge is determined by the contrast between the local gradient and the average gradient of the adjacent pixels	LIVE database PCC = 0.9613, SROCC = 0.9514
Gabarda and Cristóbal [37]	Blur and noise	The method is based on measuring the variance of the expected entropy of a given image on a set of predefined directions. Entropy can be calculated on a local basis using a spatial/spatial-frequency distribution as an approximation for a probability density function. A pixel-by-pixel entropy value is calculated. The anisotropy measure is used as an index to assess IQ	–
Brandao and Queluz [10]	Quantization noise	Based on natural scene statistics of the discrete cosine transform coefficients, modeled by a Laplace probability density function. The resulting coefficient distributions are then used for estimating the local error due to lossy encoding. Local error estimates are also perceptually weighted, using a perceptual model by [121]	LIVE database RMS = 7.439, PCC = 0.973, SROCC = 0.978
Choi et al. [18]	Blur and noise	Blur is estimated by the difference between the intensity of the current pixel and the average of neighbor pixels, the difference is normalized by the average	LIVE database PCC = −0.91
Ciancio et al. [19]	Blur	An over-complete wavelet transform of the image is computed. Coefficients of sub-bands with the same orientation are expected to be located in similar positions. Coefficients are classified as coherent or incoherent, and the blur estimation is calculated as the mean of the standard deviations of the image components associated to the incoherent coefficients	6,580 images with simulated and real blur PCC = 0.5–0.75
Ferzli and Karam [34]	Blur	The metric integrates the concept of just noticeable blur into a probability summation model which takes into account the response of the HVS to sharpness at different contrast levels	LIVE database PCC = 0.932, SROCC = 0.936
Suthaharan [102]	Blockiness	Defined in the frequency domain. Considers a JPEG compressed image (CE) as a combination of primary edges (PE), undistorted image edges (UE) and blocking artifacts (distorted image edges and block edges). The method estimates PE and UE and then filters them out from CE to obtain an estimate for blockiness	Scores for 7 images in the LIVE Database PCC = 083–0.99
Cohen and Yitzhaky [20]	Blur and noise	Evaluates noise impact in spatial and frequency domain and estimates blur in the frequency domain. The common statistical properties of power spectra of natural images are used to enhance the distortion effects. The bending point location of the modified image spectrum (smoothed power spectrum multiplied by the squared spatial frequency) is used to define an index that measures noise and blur impacts	–
Zhu and Milanfar [136]	Blur and noise	The metric Q is based on singular value decomposition of local image gradient matrix, and provides a quantitative measure of true image content (i.e., sharpness and contrast as manifested in visually salient geometric features such as edges,) in the presence of noise and other disturbances	–
Chen and Bloom [15]	Blockiness	For a given image, the absolute difference between horizontally adjacent pixels is computed, normalized, and averaged along each column. A one-dimensional discrete Fourier transform is thereafter employed and a vertical blockiness measure is derived. A horizontal blockiness measure is computed similarly. Finally, a blockiness measure for the given image is formulated by pooling those two directional blockiness measures	LIVE database PCC = 0.9628, SROCC = 0.9468

Table 2 continued

NR method	Artifacts	Brief description	Performance
Saad et al. [87]	Distortion generic	BLind Image Integrity Notator using DCT Statistics (BLIINDS) index: it is based on predicting image quality based on observing the statistics of local discrete cosine transform (DCT) coefficients. The probabilistic model is trained on a subset of the LIVE data. Four features are extracted from the DCT domain and are applied to local image patched at two spatial scales. Multivariate Gaussian distribution and the multivariate Laplacian distribution are considered	LIVE database SROCC = 0.7996
Liu et al. [63]	JPEG and JPEG2000	It is a Neural Network-based approach. A feed-forward NN is employed to operate on the feature vector (blockiness and blur) extracted from JPEG/JPEG2000 images	LIVE database, For JPEG PCC = 0.9623, RMSE = 0.109 For JPEG2000 PCC = 0.930, RMSE = 0.139
Chen and Bovik [17]	Blur	Natural scenes statistics models are combined with multi-resolution decomposition methods to extract reliable features. The algorithm is composed of three steps: (i) a probabilistic support vector machine is applied as a rough image quality evaluator; (ii) the detail image is used to refine the blur measurements; (iii) the blur information is pooled to predict the blur quality of images	LIVE database SROCC = 0.9352
Moorthy and Bovik [70]	Distortion generic	The Distortion Identification-based Image Verity and INtegrity Evaluation (DIIVINE) index is based on a 2-stage framework involving distortion identification followed by distortion-specific quality assessment. Assumes that natural scenes possess statistical properties which are altered in the presence of distortion	LIVE and TID2008 databases SROCC = 0.916 (LIVE), 0.889 (TID2008) PCC = 0.917 (LIVE)
Tang and Kapoor [42]	Distortion generic	The method uses a set of low-level image features in a machine learning framework to learn a mapping from these features to subjective image quality scores. Features are derived from natural image statistics, texture features and blur/noise estimation	LIVE database, PCC = 0.89
Gabarda and Cristobal [38]	Gaussian noise and Gaussian blur	The von Mises distribution of the image information is evaluated. Assuming that the concentration parameter decreases exponentially with increasing the amount of degradation, it can be used as an image quality assessment index	TID2008 database, For Gaussian noise: PCC = 0.8052, SROCC = 0.8083 For Gaussian blur: PCC = 0.9600, SROCC = 1.0000
Mittal et al. [67]	Distortion generic	The Blind/Referenceless Image Spatial QUality Evaluator (BRISQUE) operates in the spatial domain and is based on natural scene statistics. No distortion-specific features such as ringing, blur or blocking are modeled. The algorithm quantifies the naturalness in the image due to the presence of distortions. Machine learning-based approach requires training on database of human-rated distorted images	LIVE database SROCC = 0.9395, PCC = 0.9424
Ye and Doermann [128]	Distortion generic	Approach based on visual codebooks. A visual codebook consisting of Gabor-filter-based local features extracted from local image patches is used to capture complex statistics of a natural image. The codebook encodes statistics by quantizing the feature space and accumulating histograms of patch appearances	LIVE Database, PCC = 0.8955, SROCC = 0.8954
Saad et al. [88]	Distortion generic	BLIINDS-II: it uses a Bayesian approach to predict quality scores after a set of features is extracted from an image. The extracted features are based on the Natural Scene Statistics (NSS) model of the image DCT coefficients. Features are extracted over three scales. The probabilistic model is trained on a subset of the LIVE data	LIVE and TID2008 databases SROCC = 0.9306 (LIVE), 0.8442 (TID2008) PCC = 0.9302 (LIVE)
Mittal et al. [68]	Distortion generic	The natural image quality evaluator (NIQE) is based on the construction of a quality aware collection of statistical features based on a space domain natural scene statistic model; without training on human-rated distorted images	LIVE database, PCC = 0.9147, SROCC = 0.9135

degree of visual degradation of image data that are transmitted through communication networks. In image transmission, the features must be coded and transmitted with the image data on the same channel or through ancillary channels. The receiver computes the same features on the received image to verify if the original image has been corrupted during transmission. RR metrics can be also used in the image acquisition or processing scenarios where the acquisition device and the processing module can be assimilated to a transmission channel prone to errors and distortions. In these scenarios, the source and destination images can be of different nature, but notwithstanding this, a measure of image fidelity is often required. To this end, in many image domains, it is common to acquire known targets (e.g. patches of colors or objects) on which to compute the features to be evaluated and compared. Similarly to the transmission scenario, the features are computed before and after the acquisition or processing of the image to verify if and to what extent any quality degradation occurred. RR methods, in general, extract content-based or distortion-based features. Compared with FR and NR, few RR methods are available in the literature. In Table 3, a brief summary of RR methods is presented.

3.2 Indirect quality evaluation

The aforementioned IQ approaches assess the quality by taking into account the properties of the images themselves in the form of their pixels or feature values.

Image quality can also be indirectly assessed quantifying the performance of an image-based task performed by a domain expert and/or by a computational system. For example, in the framework of medical imaging, an image is of good quality if the resulting diagnosis is correct. In a biometrics system, an image of a face may be considered of good quality if the person can be reliably recognized, in an optical character recognition (OCR) system a scanned document is a good quality if all the words can be correctly interpreted. The European Commission has proposed in 1999 an image quality standard for Computed Tomography images [31]. In this standard, only two quality levels are considered: (1) Reproduction: details of anatomical structures are visible but not necessarily clearly defined; and (2) Visually sharp reproduction: Anatomical details are clearly defined. Visual sharp reproduction does not affect the quality of the diagnosis. The quality evaluation could be done by processing each image and assesses the fulfillment of the constraints and requirements of the task [65]. This can be done manually by domain experts and/or automatically by a computational system. In the case of the face-based biometric system, the quality evaluation could be done by a face recognition algorithm that processes and evaluates each image.

Regardless of the approach used (manual or automatic), by comparing the predictions with the known correct responses, several evaluation measures can be derived from an estimate of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) responses. Two common measures are sensitivity and specificity. Sensitivity denotes how well the expert or the system detects positives and is defined as TP/(TP + FN). Specificity quantifies how well false alarms are avoided, and it is defined as TN/(FP + TN).

Indirect quality assessment can be carried out also by assessing the performance of the imaging/rendering devices. Using suitable sets of images and one or more direct methods (both objective and subjective), it is possible to assess the quality of the imaging and rendering procedures. In this case, IQ is related to some measurable features of imaging/rendering devices, such as spatial resolution, color depth, etc…These features can be quantitatively assessed using standard targets (e.g. X-Rite ColorChecker® Classic [127], or the ISO 12233 Chart Data [50] and ad-hoc designed software tools (e.g. [46]). However, these measures alone are not sufficient to fully assess IQ. The camera phone image quality (CPIQ) Initiative of the International Imaging Industry Association (I3A) uses both objective and subjective characterization procedures [44].

3.3 IQA validation

Despite the time required to perform the test in a carefully controlled environment, subjective tests are at the base of objective quality metrics benchmarking. In fact, any objective metric must be validated with respect to user judgments. If the perceived quality of an image is not similarly detected by the metric, that metric must be discarded.

Given a reference dataset, objective and subjective results can be compared through different performance measures. Typical measures of performance are related to the prediction accuracy, the prediction monotonicity and the prediction consistency with respect to the subjective assessments. Some of these measures are listed below:

- The Pearson correlation coefficient (PCC) is the linear correlation coefficient between the predicted quality and the subjective scores. It measures the prediction accuracy of a metric. The PCC is computed after the objective quality metrics has been fitted to the subjective quality scores using non-linear regression functions. The correlation coefficient ranges from minus one to one. An absolute value of one implies a perfect correlation, while a value of zero implies that there is no correlation.
- The Spearman rank order correlation coefficient (SROCC) is the correlation coefficient between the predicted scores and the subjective scores. It measures the prediction

Table 3 Reduced Reference methods

RR method	Features	Brief description	Performance
Saha and Vemuri [90]	Features describing aliasing and blockiness effects	The active regions of an image (defined as those with strong edges and textures) are quantified. The metric is based on the wavelet coefficients from the different sub-band coding schemes and is used to predict the PSNR of compressed images	*R*-squared value = 0.9934
Kusuma and Zepernick [57]	Features describing blocking and blurring artifacts	Hybrid IQ metric. The importance of blocking effect is computed using the Wang and Bovik method [116], and the importance of blurring is measured using Marziliano's method [66]	–
Wang and Simoncelli [115]	Features describing the histograms of wavelet coefficients	Based on a natural image statistic model in the wavelet transform domain. The marginal distribution of the wavelet coefficients within a given sub-band changes in different ways for different types of image distortions. Uses an information distance measure between probability distributions to quantify such changes. No specific distortion model is assumed	LIVE database PCC = 0.9695 (JPG), 0.8889 (noise), 0.8872 (blur), 0.9353 (JPG2K) SROCC = 0.8908 (JPG), 0.8639 (noise), 0.9145 (blur), 0.9298 (JPG2K) OR = 0.0341 (JPG), 0.1793 (noise), 0.1172 (blur), 0.069 (JPG2K)
Carnec et al. [13]	Visual features similar to those used by the HVS: orientation, length, width and magnitude of the contrast at the characteristic point	Implements an operating and organizational model of the HVS, including important stages of vision (perceptual color space, CSF, psychophysical sub-band decomposition, masking effect modeling). The criterion extracts structural information from the representation of images in a perceptual space. Extracted features are stored in a reduced description which is generic, as it is not designed for specific types of distortions	IVC, LIVE and Toyama databases PCC = 0.913–0.972 ROCC = 0.909–0.953 OR = 0.02–0.05
Li and Wang [61]	Statistical features extracted from a divisive normalization-based image representation	Inspired by the success of the divisive normalization transform as a perceptually and statistically motivated image representation. Each coefficient of the transform is normalized (divided) by the energy of a cluster of neighboring coefficients. It is a general-purpose method, no assumption is made about the types of distortions present in the images	LIVE database PCC = 0.9162 SROCC = 0.9279 OR = 0.1079
Soundararajan and Bovik [100]	Entropy of Wavelet coefficients	Reduced Reference Entropic Differencing (RRED) index: the algorithm measures the changes in suitably weighted entropies between the reference and distorted images in the wavelet domain	LIVE and TID2008 databases SROCC = 0.8606 (LIVE) 0.824 (TID2008)
Rehman and Wang [84]	Statistical features extracted from a multiscale multi-orientation divisive normalization transform	The method is based on natural image statistics modeling and develops a distortion measure by following the philosophy in the construction of SSIM	SROCC = 0.9129 (LIVE), 0.8154 (IVC), 0.7210 (TID2008), 0.8003 (Toyama), 0.8527 (CSIQ), 0.7301 (A57)

monotonicity of a metric, i.e. the degree to which the predictions of a metric agree with the relative magnitudes of the subjective ratings. The range of Spearman Correlation is from minus one to one with the same significance as the Pearson correlation coefficient.

– The outlier ratio (OR) is defined as the percentage of the number of predictions outside the range of ± 2 times the standard deviations of the subjective results. It measures the degree to which the metric maintains the prediction accuracy (i.e. prediction consistency). The range of the ratio is from zero to one, with zero indicating the absence of outliers.
– The Root Mean Square Error (RMSE). The RMSE lower bound is zero indicating a perfect correspondence.

Table 4 Image quality databases

Database	Brief description
LIVE Sheikh et al. [97]	29 reference images, 779 test images, 20-29 observers/image. Distortion types: JPEG compresses images (169 images), JPEG2000 compressed images (175 images), Gaussian blur (145 images), White noise (145 images), Bit errors in JPEG2000 bit stream (145 images)
MICT Sazzad et al. [91]	14 reference images, 168 test images, 16 observers/image. Distortion types: JPEG and JPEG2000
IVC Callet and Autrusseau [12]	10 reference images, 235 test images, 15 observers/image. Distortion types: JPEG, JPEG2000, LAR coding, Blurring
A57 Chandler and Hemami [14]	Three original images and 54 distorted images (3 images, 6 distortion types, 3 contrasts). Distortion types: additive Gaussian white noise, Baseline JPEG compression, JPEG-2000 compression using different settings, Gaussian blurring, quantization of the LH sub-bands of a 5-level DWT of the image
Toyama's Database [110]	182 images of 768 × 512 pixels. Out of all, 14 were original images (24 bit/pixel RGB) in each group. The rest of the images were JPEG and JPEG2000 coded images (i.e. 84 compressed images for each type of distortion). Six quality scales and six compression ratios were, respectively, selected for the JPEG and JPEG2000 encoders
TID2008 Ponomarenko et al. [81]	25 reference images, 1,700 test images, observers/image. Distortion types: noise (Gaussian, spatially correlated, masked, high frequency, impulse, quantization, pattern), Gaussian blur, compression and transmission (JPEG and JPEG2000), blocking, intensity shift and contrast change
CSIQ Larson and Chandler [59]	30 original images, each is distorted using six different types of distortions at four to five different levels of distortion. Distortion types: JPEG compression, JPEG-2000 compression, global contrast decrements, additive pink Gaussian noise, and Gaussian blurring. This results in 866 distorted versions of original images. 5000 subjective ratings from 35 different observers
LIVE multi-distortion Jayaraman and Bovik [28]	15 reference images and 405 multiply distorted images. Four levels of blur, JPEG compression and noise are considered. The multiple distorted images consist of blur followed by JPEG and blur followed by noise. The scores are collected from 37 observers

– The *R*-squared value reflects the proportion of variation explained by a regression curve. The range of this index is from zero to one with one indicating a perfect prediction.

Different standard databases are available to test the algorithms' performance with respect to the human subjective judgments. Among the most frequently used we can cite: LIVE [97], MICT [91], TID2008 [81], IVC [12], the Toyama's database [110], CSIQ [59] and the A57 [14]. Table 4 describes these databases.

When available, several of these correlation measures and the reference databases are indicated in the last column of Tables 1, 2 and 3.

Although the performance measures above described may give an idea of how well a given metric correlates with human perception, it may be misleading to use only them to select a metric to be used in a given domain for a given task. Considering, for example, the SROCC of different metrics on the same dataset, we can rank methods. However, as it can be seen from Tables 1, 2 and 3 not all the metrics are validated on the same database.

4 Applying IQA in a production workflow

After reviewing the different types and criteria used to classify the available metrics, we pose at this point the question of how to apply these metrics. Of course, the best general purpose metric does not exist. In general, different metrics may be required at different stages of the image production workflow chain. Even if a given task and scenario would require specific metrics, we can sketch some general guidelines.

In Fig. 4 a generic image workflow chain is shown. It starts with the source data (e.g. natural scene, phenomenon, measured values, etc…) to be captured and coded by a digital image. The source can be specific of a narrow domain (e.g. resonance image where the images show a limited variability in content and technical features of the imaging device) or

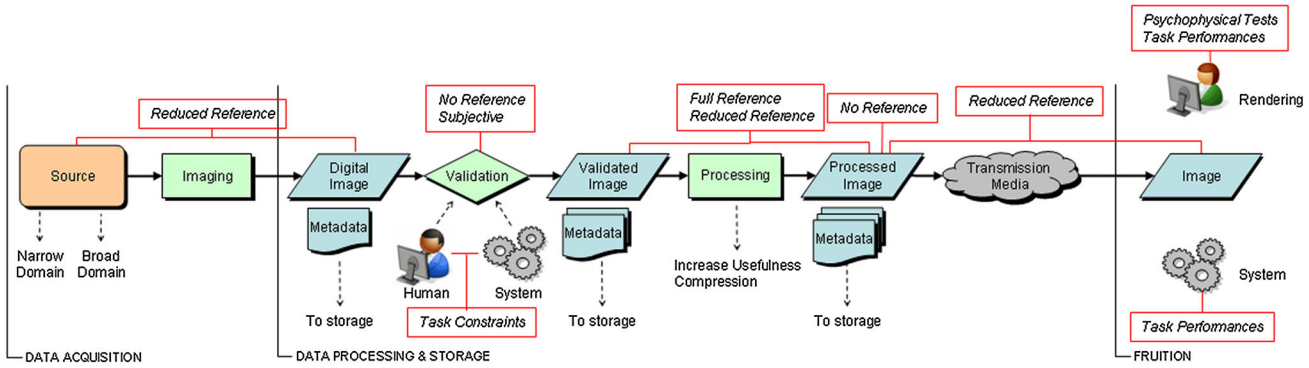

Fig. 4 Relationship between the image production workflow chain and the image quality assessment approaches

broad domain (e.g. consumer photos acquired with different digital camera, different environment conditions and with heterogeneous content). In the first case, we can reasonably expect that all the images are affected by similar distortions, while in the second case the dis-homogeneity of the images does not allow to make this assumption. Therefore, in the narrow domain, it could be simpler to find/design a pool of metrics that assess the quality of the acquired images.

The imaging block in Fig. 4 broadly refers to any imaging device, hardware or software, that transforms the source into a digital image. At this stage, quality evaluation is possible only using RR methods since the source and the digital image cannot be directly compared (for example real scene vs digital image).

The RR metric should be chosen among those in Table 3 that have been tested on datasets of images similar to the current ones. It is important to check that the datasets used to validate the RR metric chosen include images with similar contents, type, range and distribution of distortions to be detected/estimated within the given application. If one or more metrics satisfying this constraint can be found, then the performance correlations in the last column of Table 3 can be used as indicators to choose a metric. If not, a proper database (representative of the source data) should be specifically designed; psychovisual experiments should be run on this database to obtain the subjective scores that will let us validate the RR metrics. The metric showing the best performance correlation is the one to be considered. This procedure should be applied within the production workflow chain, each time a metric has to be selected.

The digital image may go through a validation phase that is aimed to have an initial assessment of the suitability and/or quality of the image with respect to the application needs (task constraints). This phase can be performed by a visual inspection. Automatic validation can also be performed using NR metrics to assess image quality or by applying rules to check if semantic constraints (e.g. minimum image size, completeness, etc.) are satisfied. The NR metrics have to be chosen among those tailored to detect the expected acquisition distortion artifacts and using the performance measures indicated in Sect. 3.3 to choose among the candidate ones. Images that do not pass the validation phase are rejected.

If required, the image can be further processed to increase its usefulness for the task at hand (e.g. contrast enhancement or binarization) or to allow more efficient transmission and storage (e.g. compression). In particular, in the latter case, FR assessment techniques can be used since the two digital images (before and after compression) are available. The image can now be transmitted and finally used either by a human observer or by an application. At the receiver/user side, further quality evaluation can be performed. Typically this amounts to ensure that the rendering devices (displays and printers) are properly calibrated and characterized for a faithful image reproduction. This can be achieved, for example, by including color profiles into the images themselves [49] as metadata information. Other metadata information can be also included along the production pipeline. Usually this information refers to the source acquisition device, image processing steps, and image content.

In what follows we illustrate IQA within three particular workflow chains. For these chains, high level IQA policies can be sketched with reference to broad categories of IQA methods (FR, NR, RR). Finally, a real case study is discussed in greater detail. The three workflow chains refer to: high-quality image archives, biometric system and consumer collections of personal photos. These scenarios have been chosen to illustrate different declination of image quality and the different constraints and requirements that must be observed while assessing it. As real case we have chosen a printing workflow for which we suggest and actually evaluate the performance of a set of specific NR IQA methods.

4.1 Quality assessment in high-quality image archives

Figure 5 illustrates an image workflow chain aimed for the population of a generic image archive for professional users such as institutional museums, photo agencies and any entity

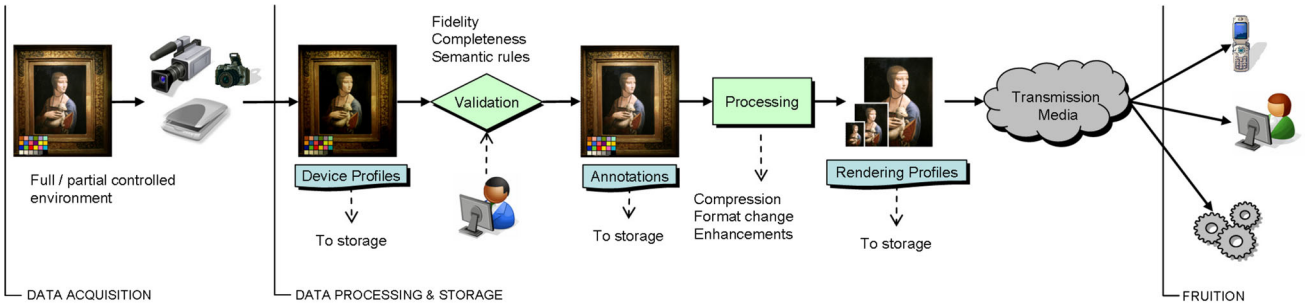

Fig. 5 Image workflow chain of a high-quality digital images archive

responsible for the management and distribution of high-quality image archives. Notwithstanding that, in this scenario, the main scope of the workflow chain is to collect images with the maximum fidelity to preserve the characteristics of the originals as much as possible. One of the major issues institutions should consider, before the digitalization process starts, is the anticipated use of their digital images [36]. Since the images can be used with different scopes, for each image several copies with different quality levels are needed. For this reason, images in a high quality archive can be classified into master, access, and thumbnail images depending on their final usage [3]. Taking into account the four categories (photograph, text, maps and graphics), some basic guidelines exist on how to generate these three groups for each of the categories [6].

In this scenario, the acquisition environment can be considered as only partially controlled. For example, in the case of an art gallery, the environment cannot be excessively tampered with to properly light the objects or move them in a better place to facilitate the acquisition procedures, paintings must be illuminated with lights that must not harm the colors or the camera cannot be freely placed in front of the objects. The acquisition is to be performed by high-end acquisition devices. Special devices can also be used to acquire large surfaces at high (or extremely high) resolution (e.g. The Google Art Project[1]). Since the fidelity of the acquired image is of paramount importance, color charts are usually used to calibrate and characterize the acquisition devices [95]. They may be also acquired along with the objects constituting reliable references for RR IQs and subsequent processing steps. In the validation phase, it is important to assess that the whole object has been completely and correctly acquired as in the case of the multi-view acquisitions of 3D objects or the surface tessellation of very large paintings. At this stage, image quality can be assessed using RR or NR methods.

Since the images collected may be distributed and used in different ways, the processing phase may include resizing, thumbnail creation, digital image format changes and compression to derive access and thumbnail images. For images that undergo processing steps, quality check is mandatory and according to the type of processing FR, RR or NR methods can be applied. With respect to image resizing, in [85], the problem of image quality for super-resolution images is addressed. In fact, super-resolution introduces blurring, aliasing, and added noise to the processed image. The quality assessment approach in [85] uses the SSIM metric to compare an image with the corresponding super-resolution one obtained by downsampling and upsampling the original. In [130] instead, a more generic approach is used which is based on natural scene statistics. Statistical models computed on high-quality natural images are built and the departures from such models are used to quantify image quality degradations.

During fruition, the perceived image quality is affected by the rendering device and the viewing conditions. For a faithful reproduction of digital images, the rendering devices must be carefully calibrated and characterized [95]. A best practice is to employ a color management system (CMS) based on the International Color Consortium (ICC) color management model [49]. At this stage, IQA can be carried out using subjective methodologies to evaluate the perceived image quality in the fruition environments.

In this high-quality image scenario, high dynamic range (HDR) images are becoming more widely available since they allow colors to be captured at the highest fidelity. However, this poses the problem of how to visualize HDR images on standard displays designed for low dynamic range (LDR) images. Usually a tone mapping algorithm is required to obtain LDR images from HDR ones. Recently, [129] proposed an objective IQA algorithm to evaluate tone-mapped images by combining structural similarity and natural scene statistics approaches.

The metadata associated with the acquired images are of great importance in this application scenario. In [6], four types of metadata schemes that institutions can use to describe their artifacts are described: Descriptive Metadata, Administrative Metadata, Structural Metadata and Technical Metadata. Since all the stored textual data are intertwined

[1] http://www.googleartproject.com/.

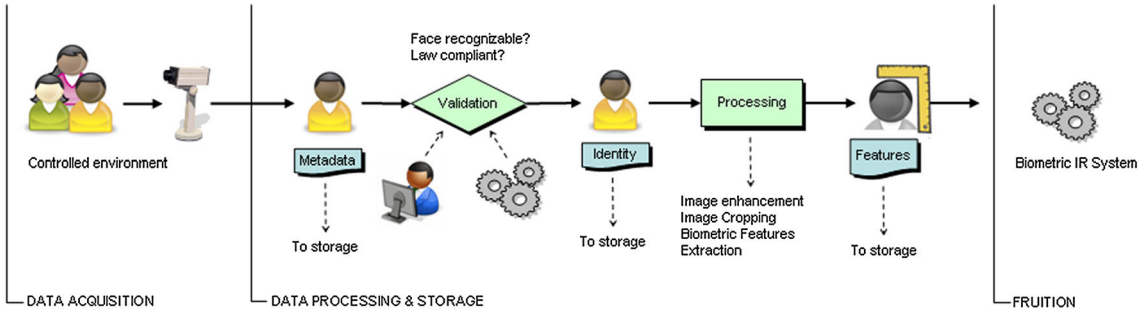

Fig. 6 Image workflow chain of a generic biometric system exploiting facial information

and can refer to the same conceptual entities in complex and complementary ways, the quality of the data stored into the database should be also taken into account with appropriate metrics (see [4,5]).

4.2 Quality assessment in a biometric system

In authentication and verification biometric systems, measuring the quality of biometric samples is a crucial step during the enrolment process. Biometric matching algorithms depend on quality of sample features that can be extracted from digitized samples. Good-quality biometric samples obviously increase the performance of the matching process. There are different causes that influence the quality of the biometric sample image such as environment, acquisition device, operator, and user cooperation. Maintaining a consistent level of sample's quality allows a more robust discrimination among samples during authentication or verification.

A general workflow chain for a biometric system for digital passports is shown in Fig. 6. Since the application needs and constraints are strictly defined, very low variability in the workflow chain is admitted. The source images are usually taken in a controlled environment. Lights and subject pose are managed to reduce the variability in the acquisition at minimum and to focus on the face of the subject. The acquisition device is usually a digital camera. It must have enough spatial resolution to make it possible to record all the relevant facial information, without visible artifacts. The latter can be detected with NR and RR methods. Furthermore, acquired images pass through a validation phase that is aimed to discard those images that do not possess the requirements adopted in digital passports. The International Civil Aviation Organization (ICAO) based its requirements [45] on the the ANSI INCITS 385-2004 standard [2], later to become an ISO/IEC IS 19794-5 standard [51]. Among the requirements for a good quality facial image are: be in sharp focus and clear; show skin tones naturally; have an appropriate brightness and contrast; be color neutral; show open eyes and clearly visible (no hair across eyes); be taken with a plain light colored background; be taken with uniform lighting; no shadows or flash reflections on the face; no red-eye effect. For the full set of 23 features that must be taken into account for a compliant face image, see [51].

The validation phase can be carried out by a domain expert or by an automatic procedure. For example, the FaceQM tool [131] is able to automatically evaluate 15 out of 23 ANSI 2004 image requirements. A frontal token image can be created to enable a facial recognition algorithm to operate more efficiently. The token image must satisfy specific requisites. The validated images can then be stored in a database with biometrical features (eye location, eye to mouth distance) or can also be extracted and stored during this stage. In the case of face-based authentication system, the quality of the acquired images can be also indirectly assessed using False Accept Rate (FAR) and False Rejection Rate (FRR) indices. From the above observations, we can see that in this scenario, the task constraints are more important with respect to the quality approaches described in Sect. 3.

4.3 Quality assessment in consumer collections of personal photos

In personal photo collections, more than in the other scenarios, the definition of what is intended by image quality is rather fuzzy (see Fig. 7). In high-quality catalogues and in biometric systems, the source and use of the images are well defined. On the contrary, in personal collections, these factors are not well defined. Consequently, the variability in the image production chain is much larger. The subject and the scene depicted in the images are varied and unconstrained, images can be acquired using a variety of devices, and the resulting images can be used in different ways (stored, printed or shared) that are not necessarily known in advance. In summary, personal photo collections belong by definition to the broad image domain and thus it is very difficult to define a general IQA.

Judging the quality of personal photos only on the basis of the presence of low level distortions, such as noise or blur, may be not sufficient since in this scenario there are no con-

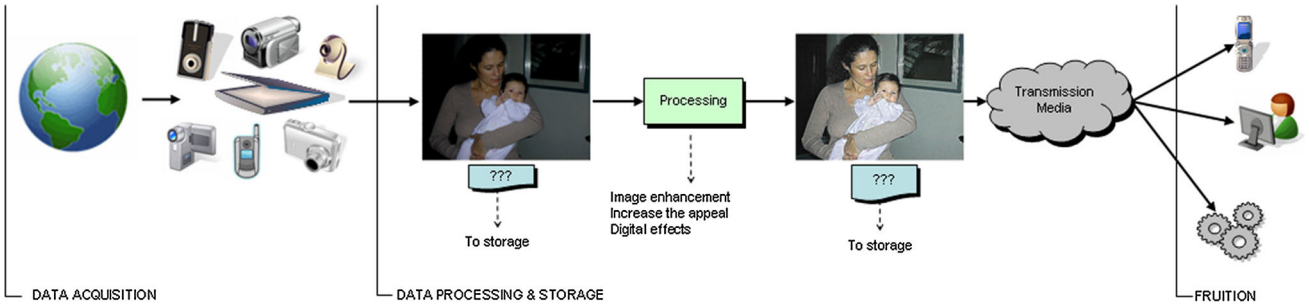

Fig. 7 Image workflow chain of a personal photo collection

straints on how the images have been acquired or why they have been acquired in such a way. Personal photos are usually taken to record special events and often users pay little attention to the acquisition conditions: images taken with very low light are usually very noisy but they can be considered acceptable if they depict properly the object of interest. An out of focus image may be considered a good image if it was taken with amusing or appealing intentions in mind or if it preserves a cherished memory of some event [56]. Compositional rules such as the *Rule of Thirds* or the *Equalized visual weights* [54] are seldom observed by occasional photographers. Thus, the validation phase is mainly subjective and may vary from photo to photo. Among the defects that can be automatically assessed on consumer photos, there is the red-eye detection. Red eye is one of the most common defects that even occasional photographers pay attention to and is also a recoverable one, i.e., a computational procedure can be applied to remove it from the images [39]. Gross acquisition defects can be detected with NR IQ metrics and could be used to filter out bad images. Other possible criteria to automatically reject images are to discard near duplicated ones. It is not uncommon to take many photos in sequence (usually during some relevant occasions such as parties) and an automatic procedure can be devised to detect them [133].

If we want to take into account also aesthetics, several approaches can be used to evaluate it. For example, in [30], low level features are used to classify images into aesthetically pleasing and displeasing. Exposure of light, colorfulness, saturation, hue, the rule of thirds, familiarity, size, aspect ratio and low depth of field are some of the features exploited. At a less tangible level, we can even consider an image from the point of view of color preference and color harmony [92,99] or how to detect combinations of pleasing colors or sets of colors that inspire particular moods [26].

Some of the images passing the validation phase may undergo an enhancement process. Apart from the usual geometric processing such as scaling, cropping and rotation, users may desire the images to look good or even funny or conspicuous disregarding the fidelity with respect to the scene depicted. The perceptual or subjective impression is considered more important than the objective quality. Consequently, the processing is mainly aimed to enhance the image appeal, even if the processing steps may introduce defects or distortions such as halos or unnaturally high color saturation. The processing procedures may also include the application of digital effects to make the image more attractive for potential viewers. In this scenario, images are mostly stored in JPEG format since this is the most common output of mid- low-end cameras and it is a suitable format for image sharing.

This dependence requires the design of IQ metrics that take into account different distortions simultaneously.

4.4 A case study: IQA in a printing workflow

In this section, we describe our experience in IQA within a printing workflow chain for a real case study. The case study originated from an investigation whose aim was to integrate IQA metrics within one Océ-Canon [2] printing workflow chain. Specifically, the goal was to automatically classify digital images into three classes: high quality ones, that can be directly printed; low quality ones, that do not deserve printing; and medium quality images, where the printing decision would be taken by a manual operator. Without an automatic procedure, images are either all printed (with a waste of ink, paper and materials), or subjectively evaluated by an expert (with a considerable wasting of time). In Fig. 8, a representative flowchart of this printing chain is depicted. A quality validation module is introduced that, using NR metrics, permits to classify the images as high, medium or low quality. High-quality images will be directly printed. Low quality ones will be rejected, while the rest of them will be forwarded to the human judgment. After this step of subjective quality evaluation, some images could have been evaluated good enough to be printed, others will be discarded, while some of them can be sent to a processing module where image enhancement can be performed using basic software or professional ones. Among all the possible distortions, the company was mainly interested on JPEG and noise artifacts and the combined effect of noise followed by JPEG compression.

[2] http://www.oce.com.

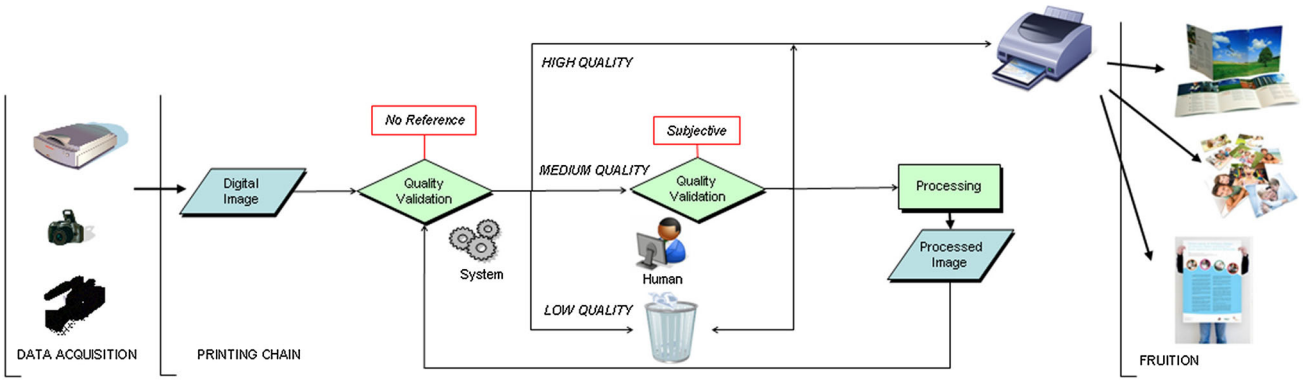

Fig. 8 IQA in the analyzed printing workflow chain

As we have previously emphasized (see Sect. 4), it is important to individuate a proper database where the performance of the NR metrics to be chosen for this specific application will be evaluated. By proper we mean a database where both type and range of distortions are representative of the real case studied. After reviewing the available databases (see Table 4), we have found that none of them is suitable for this application [22]. In general, the distortion range in the available databases varies from images of high quality to images highly corrupted, with a greater emphasis with respect to degraded images. This is due to the fact that most of these databases were generated for academic purposes. However, images considered in the real application under study present a narrower range of distortions, more concentrated in the medium and high quality. To this end, a proper database [22,23] was specifically generated to satisfy the constrains of the printing company. This database originates from 20 reference images of 886 × 591 pixels (15 × 10 cm at 150 dpi, typical printing parameters for natural photos), chosen to sample different contents both in terms of low level features (frequencies, colors) and higher ones (face, buildings, close-up, outdoor, landscape). The corresponding thumbnails are shown in Fig. 9.

Starting from these images, the whole database is composed of:

- Noise database: 200 noisy images obtained as follows: for each of the 20 reference images, we have created 10 corrupted versions with: 1, 2, 3, 4, 5, 6, 8, 10, 12 and 14 gray levels of standard deviation on the luminance channel.
- JPEG database: 180 JPEG compressed images generated using the Matlab imwrite function. As the Q-factor depends on the specific JPEG compression algorithm used, we have adopted the bit per pixel (bpp) Ratio (bppR) with respect to a reference, finding iteratively the Q-factors that better match the corresponding bppR values. As reference we have adopted the $Q = 100$ compressed image, where the compression is mainly due to the sub sampling of the chroma channels and to lossless algorithms. For each of the 20 original images, we have created 9 compressed versions with the following bppR: 1 ($Q = 100$), 0.707, 0.5, 0.25, 0.177, 0.125, 0.105, 0.088, 0.0625.
- MD database: 800 Multiply Distorted images generated as follows: each of the 200 noisy images were further processed by 4 different levels of JPEG compression, corresponding to Q factor values of 100, 50, 30, and 10.

For collecting the subjective data, in terms of Mean Opinion Scores (MOS), a Single Stimulus method was adopted, where all the images are individually shown. Observers were asked to provide their perception of quality on a continuous linear scale that was divided into five regions, marked with adjectives (*Bad*, *Poor*, *Fair*, *Good*, and *Excellent*). The scale was then converted into 1–100 linearly. The experiments were performed following the recommendations in ITU [52].

In what follows we report in details the results on the JPEG database, and then we summarize the results obtained for the noise database, and the MD database. For what concerns the objective data that have to be correlated with the subjective evaluations, working on JPEG, the following NR metrics are here considered (M1–M4 for single distortion, M5, and M6 general purpose).

- M1: by Wu and Yuen [126]
- M2: by Wang et al. [116]
- M3: by Muijs and Kirenko [72]
- M4: by Chen and Bloom [15]
- M5: BRISQUE metric, by Mittal et al. [67]
- M6: NIQE metric, by Mittal et al. [68]

The metrics and the subjective scores have been correlated using a logistic function. Metrics that correlate highly with

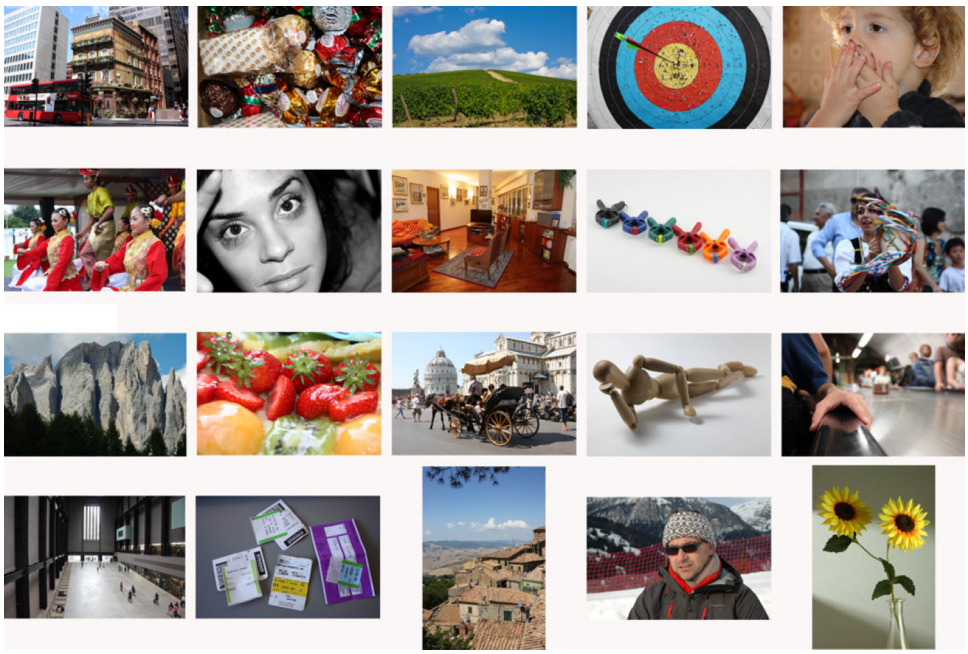

Fig. 9 The 20 reference images of the database

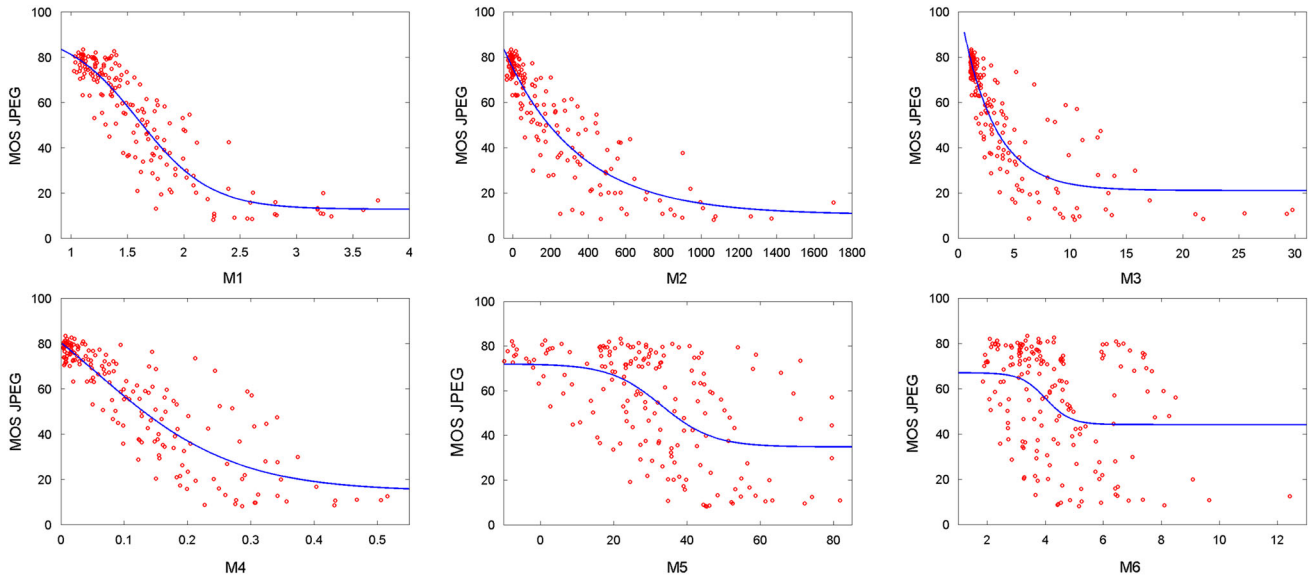

Fig. 10 Logistic regression for NR metrics and JPEG database. *First row* M1, M2, and M3, *second row* M4, M5 and M6

human ratings typically yield high Pearson and Spearman correlation coefficients (greater than 0.9).

In Fig. 10, we report the logistic correlations between each of the M1–M6 metric and the subjective scores. In Table 5, the corresponding correlation coefficients are shown.

Keeping in mind our classification task within high, medium, and low quality classes, we have grouped the MOS scores so that images evaluated as *Bad* and *Poor* correspond to our low quality class, images evaluated as *Excellent* and *Good* correspond to high quality one, while images *Fair* are assigned to class medium. These groups correspond to the ground truth of our classification problem. The predicted classes are obtained applying thresholds to the regression curve between an NR metric and the subjective scores.

As example we here report in detail the classification obtained using the M2 metric, as it shows the highest correlation coefficients (see Table 5). In Fig. 11 (left), we show the MOS scores corresponding to images evaluated, respectively, as high, medium and low quality ones. The predicted classes obtained by thresholding this regression curve are shown in Fig. 11 (right). The performance of this classification is reported in Table 6 in terms of confusion matrix.

Table 5 Pearson and Spearman correlation coefficients for NR JPEG-blockiness (M1–M4) and general purpose metrics (M5, M6) on the JPEG database

Correlation	M1	M2	M3	M4	M5	M6
PCC	0.9028	0.9059	0.8789	0.8685	0.5747	0.3593
SROCC	0.8662	0.8922	0.8714	0.8689	0.5387	0.3534

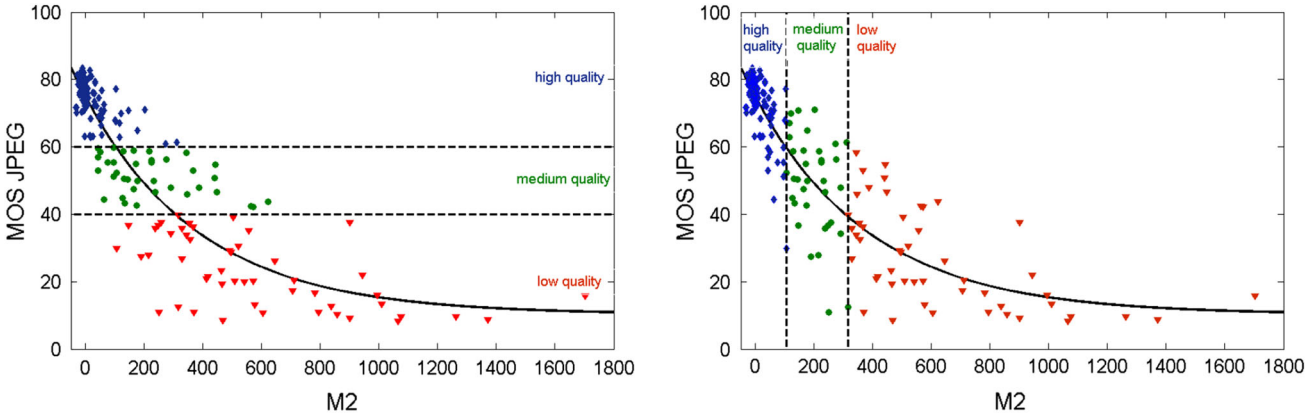

Fig. 11 MOS scores versus M2 metric. *Left* ground truth with respect to the three classes, high, medium, and low. *Right* the corresponding predicted classes obtained thresholding the regression curve

Table 6 Confusion matrix obtained using the regression curve of M2 metric and JPEG database

Class	Predicted		
real	Low	Medium	High
Low	85	8	0
Medium	9	18	10
High	0	10	40

Error = 20 %

Table 7 Classification errors for NR JPEG-blockiness (M1–M4) and general purpose metrics (M5, M6) on the JPEG database

Metric	M1	M2	M3	M4	M5	M6
Classification error (%)	22	20	19	23	49	57

Each column of the matrix represents the instances in a predicted class, while each row represents the instances in an actual class. All correct predictions are located in the diagonal of the table. All the non-zero elements outside the diagonal represent misclassifications. The performance error is obtained as the ratio between the misclassified elements and the total number of images. Summarizing, for the M2 metric and JPEG database, the error performance for the classification task is 20 %. The classification results considering all the 6 NR metrics above cited, summarized in terms of classification errors, are reported in Table 7.

In conclusion, for our classification task, distortion specific metrics (M1–M4) perform better than general purpose ones (M5, M6) as also suggested by the correlation coefficients (see Table 5). Among distortion specific metrics, taking into account the values of PCC and SROCC and the classification errors, metrics M2, and M3 are the most suitable metrics to be adopted within the considered workflow (see Fig. 8).

For what concerns the noise database, we here consider as distortion specific metric only the NR metric by Immerkaer [47] (hereafter called M7), as this metric highly correlates with the subjective data in case of Gaussian noise. The metric implementation by Foi [35] is used in what follows. As general purpose metrics, we consider the same M5, and M6 applied also in the case of JPEG compression.

In Fig. 12, the logistic regressions of these NR metrics with the subjective scores of the noise database are reported. The corresponding PCCs and SROCCs and classification errors are summarized in Table 8.

Finally for what concerns the MD database, we here consider the metric M7 for noise and M2 for specific JPEG distortion as it is the best one among those considered for the JPEG database. Also the two general purpose metrics M5 and M6 are taken into account. In Fig. 13, the subjective scores of the MD database are plotted versus the M2, M7, M5, and M6 metrics, respectively. Note that in case of M7, it was not possible to find a reasonable logistic regression. From these plots, we observe that neither metrics specifically developed for single distortion nor general purpose ones are

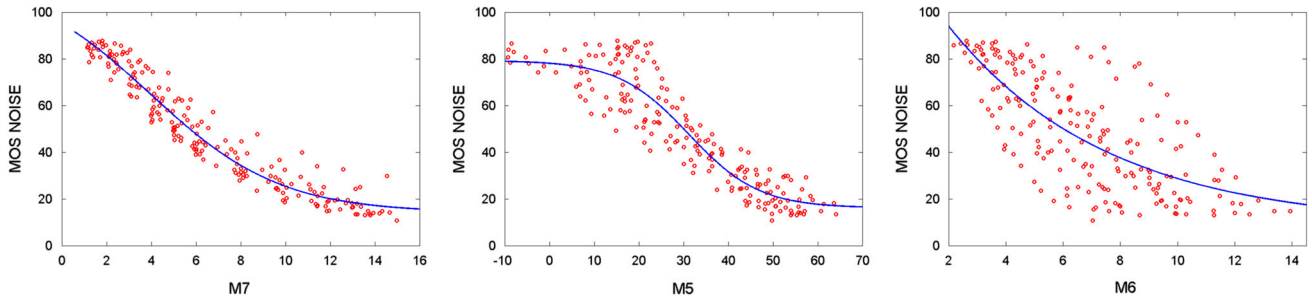

Fig. 12 Logistic regression for M7, M5, and M6 NR metrics and noise database

Table 8 Performance evaluation of the M7, M5, and M6 metrics on noise database

Metric	M7	M5	M6
SROCC	0.9660	0.9096	0.7300
PCC	0.9688	0.9262	0.7393
Classification error (%)	15	16	36

able to properly fit the subjective data in the case of multiple distortion noise-JPEG as also confirmed by the corresponding PCC and SROCC in Table 9. For this reason, we think that it has no sense to perform the classification task here proposed, for the case of MD database. This is an open issue that has to be addressed in the near future with the development of proper metrics that take into account the presence of simultaneous distortions.

Summarizing, we can sketch the following guidelines concerning the choice of NR IQA metrics:

- If the kinds of distortion corrupting the images are known, the corresponding distortion-specific metrics are the most suitable ones. Based on our experience and on the case study here presented, we propose to use:
  - the metric developed by Wang et al. [116] for JPEG artefacts;
  - the metric by Immerkaer [47] for Gaussian noise distortion;
  - the metric by Marziliano et al. [66] for Gaussian blurriness;
  - the metric by Chen and Bovik [17] in case of real blurriness (simple and complex motion blur, out-of focus, etc.);
  - the metric by Zhu and Milanfar [136] if both blurriness and noise are present.
- In case of multiple distortion or when the kinds of distortion present in the images are unknown, the following two blind metrics are suggested:
  - the Opinion-aware metric BRISQUE [67] that has been trained on the LIVE database;
  - the Opinion-unaware metric NIQE metric [68] for which no training phase is required.

5 Open issues

In this section, we consider some open issues that are being addressed at present by the IQA community.

A challenge task is how to design a general purpose NR IQ metric capable of assessing different artifacts simultaneously. It is not surprisingly that the major part of the NR IQ metrics are designed to measure only a single distortion. The few that consider two distortions simultaneously are mainly concerned with the case of noise and blur that are correlated. A first idea could be to combine different IQA metrics into a single method. However, before considering different combination strategies, the normalization problem of the single metrics should be addressed. Both the normalization and combination of multiple metrics are still open problems within the IQA community. The same issue applies if we aim to increase the performance of detecting a given artifact by combining several metrics. To cope with this problem, general purpose metrics (or universal metrics) have been proposed by [42,67,128]. Despite these methods show promising results as generic metrics, they have been tested mainly on the LIVE database and on other databases where each corrupted image is affected by a single distortion. Recently, a multi-distortion database [28] has been introduced where the multiple distortions consist of blur followed by JPEG and blur followed by noise.

Another aspect that has not been fully addressed up to now in the literature is the interference created by the two signals that compose an image: the content and the distortion. As the distortion increases, the visibility of the content decreases and, for the case of natural images, these two signals may not be clearly separated or their mix is spatially varying. This poses a problem when designing, for example, an NR metric. Even if the metric was properly designed to identify a given distortion, the content of the image can still

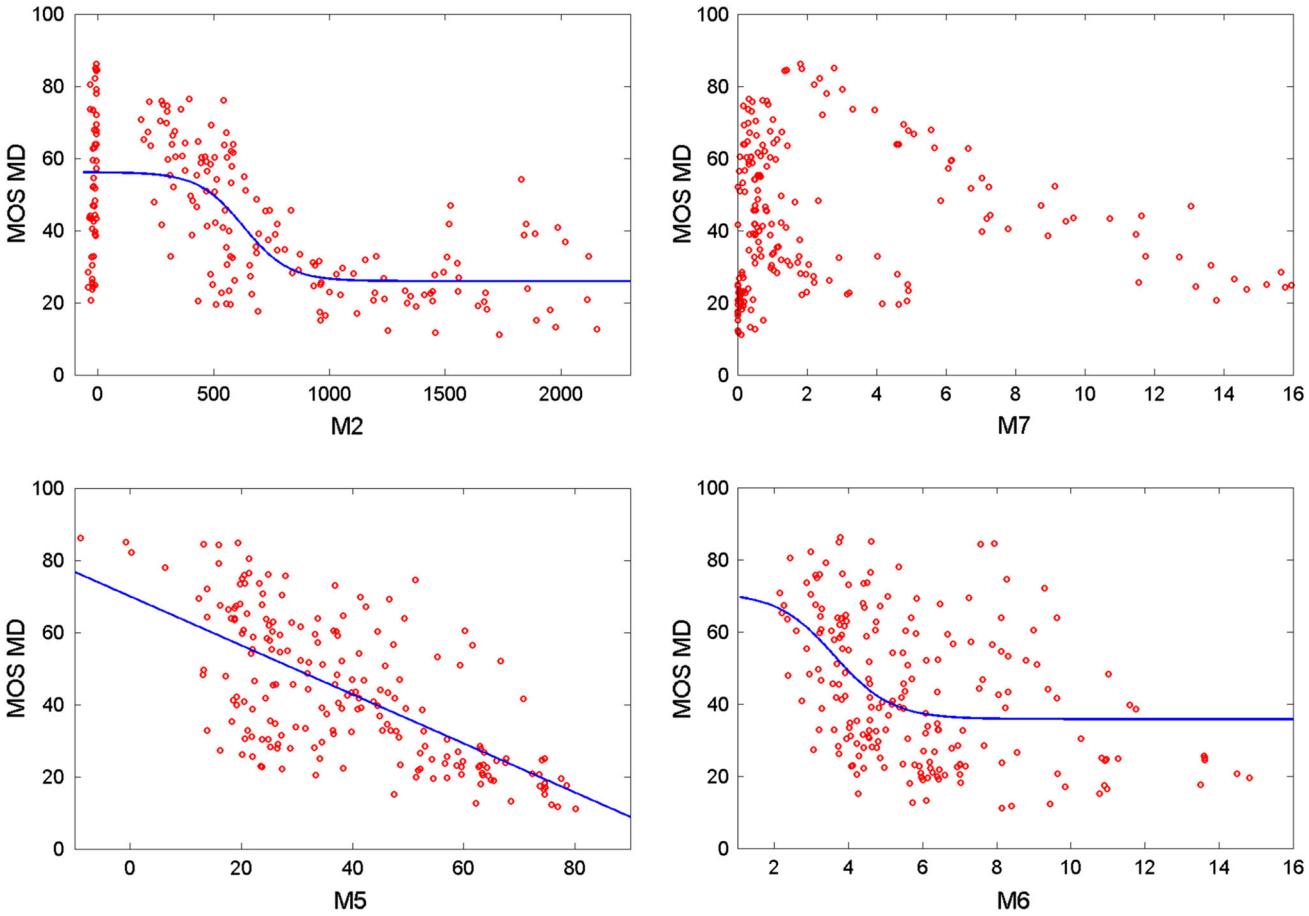

Fig. 13 Logistic regression for M2, M7, M5, and M6 NR metrics and MD database

Table 9 Performance evaluation of the M2, M7, M5, and M6 metrics on MD database

MD	M2	M7	M5	M6
SROCC	0.5575	–	0.6541	0.4371
PCC	0.6515	–	0.6583	0.4407

influence the metric estimation because the metric is blind to the content of the image and cannot distinguish between content signal and distortion signal. In fact, as we have already noted, an IQ measure computed on the overall image may not be representative of the perceived quality (see Fig. 2). A more suitable way could be to compute the IQ measure on selected regions chosen on the basis of their properties and of the application task. The selection of these regions can be done manually using interactive tools (e.g. [21]), and only in some cases could be automatically performed by applying a region annotation method like the one described by Cusano et al. [27].

To overcome the differences between the subjective and the objective quality metrics, many researchers have tried to integrate human vision cognition models within the quality metrics. Taking into account low level features of the Human Vision System (like for example luminance sensitivity, contrast sensitivity and texture masking) has led to the development of many HVS-based metrics [29,64,89,105]. On the other hand, the high-level features of the HVS have lately become of interest for researchers committed in image quality perception modeling. Among them, we can cite the visual attention mechanism. This feature makes the observer focus on a selected or salient region while ignoring other areas of the image. Many attempts have been done to integrate the visual saliency information within the quality metrics but the results are contradictory up to now [69,74]. Different issues arise while integrating visual saliency within quality metrics. One of them is the fact that the quality assessment task in itself can affect the way people look at an image [1,73]. Another important point to be considered is the type of saliency map used: different computational models can be used to generate these maps but also the ground truth maps can be considered (obtained from eye tracking experiments where the gaze direction is recorded) [62,83].

Besides the visual attention, other high level features, for example, prior information regarding the image contents,

affect the evaluation of the image quality. Subjective experiences and preferences may also influence the human assessment of image quality: it has been shown that the perceived distortions are dependent on how familiar the test person is with the observed image. For example, it is well known that one of the objects attracting most of our attention is people and especially human face. Some objective metrics have been proposed that try to include this information in a top-down way (see for example [107]).

Following numerous psychophysical studies, Larson et al. [59] assume that the HVS performs multiple strategies when determining quality. They argue that, given a single task of judging image quality, a human observer employs different strategies when confronted with different image conditions. In the high-quality regime (i.e. for evaluation high quality images), the HVS attempts to look for distortions in the presence of the image, whereas in the low-quality regime, the HVS attempts to look for image content in the presence of the distortions. These two fundamentally different strategies require two separate computational models. With this goal in mind, the authors proposed an FR method called Most Apparent Distortion. It operates using both a detection-based model and an appearance-based model. For detection, they employ a spatial domain model taking into account the contrast sensitivity, local luminance and contrast masking. For appearance, they employ a model that follows from the texture-analysis literature. The overall quality of the distorted image is computed by taking a weighted geometric mean of the detection-based and appearance-based qualities, where the weight is determined based on the amount of distortion. For highly distorted images, greater weight is given to the appearance-based quality, whereas for images containing near threshold distortions, greater weight is given to the detection-based quality. Therefore, it should be beneficial to consider the possibility of extending this two-strategy model to the case of NR and RR metrics.

Finally, video quality assessment (VQA) should be mentioned as a natural extension of IQA. Most of the contemporary video coding standards use motion compensation and block-based coding schemes for compression. As a result, the decoded video suffers from one or more of the compression artifacts, such as blockiness, blurriness, color bleeding, ringing, false edges, jagged motion, chrominance mismatch, and flickering. Transmission errors such as damaged or lost packets can further degrade the video quality. Furthermore, the pre- or post-processing stages in the video transmission system, such as domain conversion (analog to digital or vice-versa), frame rate conversion, and de-interlacing degrade the video. Therefore, the methods for evaluating video quality play a critical role in quality monitoring to maintain Quality of Service (QoS) requirements. Ad-hoc metrics have been designed for videos (e.g. [79,80,93]), as well as many IQ metrics have been extended to videos (e.g. [87,101]).

6 Conclusions

Image quality assessment is a very active and evolving research area. In this paper, we provided an overview of the state of the art of the IQA methods, putting in evidence their applicability and limitations in different application domains. It should be now evident to the readers that the selection and use of the different metrics depend on the semantic content of the image, the application task, and the particularly imaging chain applied. To design more reliable and general purpose image quality metrics, an interdisciplinary approach is the challenge for the next years. Evidence from the biological studies will help us to understand how our brain works when involved in the quality assessment task. Computational models of the visual system that account for these cognitive behaviors could be integrated within the perceptual quality metric design. Last but not least, semantic models from the image understanding community can certainly help us improve the metrics' design and performance.

Acknowledgments The authors would like to thank Fabrizio Marini for the insightful discussions on image quality and for his work in developing the no reference image quality tool. This work was partially supported by Océ-Canon.

References

1. Alers, H., Bos, L., Heynderickx, I.: How the task of evaluating image quality influences viewing behaviour. In: The Third International Workshop on Quality of Multimedia Experience (QoMEX), Belgium (2011)
2. ANSI: Face Recognition Format for Data Interchange. ANSI INCITS, pp. 385–2004. ANSI (2004)
3. Archives UN: Technical guidelines for digitizing archival materials for electronic access: creation of production master files—raster images. http://www.archives.gov/preservation/technical/guidelines.html. Accessed 09 Feb 2012
4. Batini, C., Scannapieco, M.: Data Quality: Concepts, Methodologies and Techniques. Data-Centric Systems and Applications. Springer Inc, New York (2006)
5. Batini, C., Barone, D., Cabitza, F., Ciocca, G., Marini, F., Pasi, G., Schettini, R.: Toward a unified model for information quality. In: VLDB'08 (2008)
6. BCR's CDP Digital Imaging Best Practices Working Group: BCR's CDP Digital Imaging Best Practices Version 2.0. http://mwdl.org/docs/digital-imaging-bp_2.0.pdf (2008)
7. Bhattacharya, S., Sukthankar, R., Shah, M.: A holistic approach to aesthetic enhancement of photographs. ACM Trans. Multimed. Comput. Commun. Appl. **7S**, 21:1–21:21 (2011)
8. Bovik, A., Liu, S.: Dct-domain blind measurement of blocking artifacts in dct-coded images. In: IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 3, pp. 1725–1728 (2001)
9. Bovik, A.C., Wang, Z.: Modern Image Quality Assessment. Morgan & Claypool Publishers, San Rafael (2006)
10. Brandao, T., Queluz, M.P.: No-reference image quality assessment based on dct domain statistics. Signal Process. **88**(4), 822–833 (2008)

11. Charrier, C., Lezoray, O., Lebrun, G.: Machine learning to design full-reference image quality assessment algorithm. Signal Process. Image Commun. **27**, 209–219 (2012)
12. Callet, P., Autrusseau, F.: Subjective quality assessment irccyn/iovc database. http://www.irccyn.ec-nantes.fr/ivcdb/ (2005)
13. Carnec, M., Callet, P.L., Barba, D.: Objective quality assessment of color images based on a generic perceptual reduced reference. Signal Process. Image Commun. **23**(4), 239–256 (2008)
14. Chandler, D., Hemami, S.: A57 image database. http://foulard.ece.cornell.edu/dmc27/vsnr/vsnr.html (2007)
15. Chen, C., Bloom, A.: A blind reference-free blockiness measure. In: Lecture Notes in Computer Science, vol. 6297, pp. 112–123. Springer, Berlin (2010)
16. Chen, G.H., Yang, C.L., Xie, S.L.: Gradient-based structural similarity for image quality assessment. In: 2006 IEEE International Conference on Image Processing, pp. 2929–2932(2007)
17. Chen, M.J., Bovik, A.C.: No-reference image blur assessment using multiscale gradient. EURASIP J. Image Video Process. **1**, 1–11 (2011)
18. Choi, M., Jung, J., Jeon, J.: No reference image quality assessment using blur and noise. Int. J. Comput. Sci. Eng. **2**(3), 76–80 (2009)
19. Ciancio, A., da Costa, A., da Silva, E., Said, A., Samadani, R., Obrador, P.: Objective no-reference image blur metric based on local phase coherence. Electron. Lett. **45**(23), 1162–1163 (2009)
20. Cohen, E., Yitzhaky, Y.: No-reference assessment of blur and noise impacts on image quality. Signal Image Video Process. **4**, 289–302 (2010)
21. Corchs, S., Gasparini, F., Marini, F., Schettini, R.: Image quality: a tool for no-reference assessment methods. In: Image Quality and System Performance VIII, IS&T/SPIE Electronic Imaging, SPIE, vol. 7867, pp. 78,760X (1–9) (2011)
22. Corchs, S., Gasparini, F., Schettini, R.: No reference image quality classification for jpeg distorted images. Digital Signal Process. (2014a, in press)
23. Corchs, S., Gasparini, F., Schettini, R.: Noisy images-jpeg compressed: subjective and objective image quality evaluation. In: IS&T/SPIE Electronic Imaging, International Society for Optics and Photonics, pp. 90,160V–90,160V (2014b)
24. Corner, B.R., Narayanan, R.M., Reichenbach, S.E.: Noise estimation in remote sensing imagery using data masking. Int. J. Remote Sens. **24**(4), 689–702 (2003)
25. Crosby, P.: Quality is Free. McGraw-Hill, New York (1979)
26. Csurka, G., Skaff, S., Marchesotti, L., Saunders, C.: Learning moods and emotions from color combinations. In: Proceedings of the Seventh Indian Conference on Computer Vision, Graphics and Image Processing, ICVGIP '10, pp. 298–305 (2010)
27. Cusano, C., Ciocca, G., Schettini, R.: Image annotation using SVM. In: Proceedings of Internet Imaging, SPIE, vol. 5304, pp. 330–338 (2004)
28. Jayaraman, D., Moorthy, A., Mittal, A., Bovik, A.: Objective quality assessment of multiply distorted images. In: Proceedings of the Asilomar Conference on Signals, Systems and Computers (2012)
29. Daly, S.J.: Visible differences predictor: an algorithm for the assessment of image fidelity. In: Rogowitz, B.E. (ed) Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, vol. 1666, pp. 2–15 (1992)
30. Datta, R., Joshi, D., Li, J., Wang, J.Z.: Studying aesthetics in photographic images using a computational approach. In: Proceedings of ECCV, pp. 7–13 (2006)
31. EC: Europeans guidelines on quality criteria for computed tomography. http://www.drs.dk/guidelines/ct/quality/. Accessed 09 Feb 2012
32. Engeldrum, P.: A short image quality model taxonomy. J. Imaging Sci. Technol. **48**(2), 160–165 (2004)
33. Fei, X., Xiao, L., Sun, Y., Wei, Z.: Perceptual image quality assessment based on structural similarity and visual masking. Signal Process. Image Commun. **27**(7), 772–783 (2012)
34. Ferzli, R., Karam, L.J.: A no-reference objective image sharpness metric based on the notion of just noticeable blur (jnb). IEEE Trans. Image Process. **18**(4), 717–728 (2009)
35. Foi, A.: Anisotropic nonparametric image restoration demobox. http://www.cs.tut.fi/lasip/2D/ (2006)
36. Franziska, S., Frey, J.M.R.: Digital imaging for photographic collections: foundations for technical standards. Technical report, Rochester Institute of Technology, Rochester (1999)
37. Gabarda, S., Cristóbal, G.: Blind image quality assessment through anisotropy. J. Opt. Soc. Am. A **24**(12), B42–B51 (2007)
38. Gabarda, S., Cristobal, G.: No-reference image quality assessment through the von mises distribution. J. Opt. Soc. Am. A **29**(10), 2058–2066 (2012)
39. Gasparini, F., Schettini, R.: A review of redeye detection and removal in digital images through patents. Recent Pat. Electr. Eng. **2**(1), 45–53 (2009)
40. Girod, B.: What's wrong with mean-squared error? In: Watson, A.B. (ed.) Digital Images and Human Vision, pp. 207–220. MIT Press, Cambridge (1993)
41. Gonzales, R.C., Woods, R.: Digital Image Processing. Prentice Hall, Englewood Cliffs (2008)
42. Tang, H., Joshi, N., Kapoor, A.: Learning a blind measure of perceptual image quality. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 305–312 (2011)
43. Hasler, D., Süsstrunk, S.E.: Measuring colorfulness in natural images. In: Rogowitz, B.E., Pappas, T.N. (eds.) Human Vision and Electronic Imaging VIII. SPIE, vol. 5007, pp. 87–95 (2003)
44. I3A: Fundamentals and review of considered test methods. CPIQ Initiative Phase 1 White Paper (2007)
45. ICAO-NTWG: Machine readable travel documents (MRTDs): history, interoperability, and implementation. Technical report, International civil aviation organization (2007)
46. Imatest: Digital Image Quality Testing. http://www.imatest.com (2010)
47. Immerkaer, J.: Fast noise variance estimation. Comput. Vis. Image Underst. **64**(2), 300–302 (1996)
48. ISO: Quality management and quality assurance. vocabulary (2000). iso 84021994
49. ISO: Image technology colour management—architecture, profile format and data structure? Part 1: based on icc.1:2004-10 (2005). iso 15076-1
50. ISO: ISO 12233 Chart Data. http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=33715. Accessed 09 Feb 2012
51. ISO/IEC: Biometric data interchange formats—part 5: face image data (2004). is19794-5
52. ITU: Methodology for the subjective assessment of the quality for television pictures. Technical report, ITU-R Rec. BT. 500-13 (01.12) (2012)
53. Janssen, T.: Computational Image Quality. SPIE Press, Bellingham (2001)
54. Jonas, P.: Photographic Composition Simplified. Amphoto Publisher, New York (1976)
55. Juran, J.: Juran on Planning for Quality. The Free Press, New York (1988)
56. Keelan, B.W.: Handbook of Image Quality: Characterization and Prediction. CRC Press, Boca Raton (2002)
57. Kusuma, T., Zepernick, H.J.: A reduced-reference perceptual quality metric for in-service image quality assessment. In: Joint First Workshop on Mobile Future and Symposium on Trends in Communications, 2003. SympoTIC '03, pp. 71–74 (2003)
58. Laparra, V., Muoz, J., Malo, J.: Divisive normalization image quality metric revisited. J. Opt. Soc. Am. A **27**(4), 852–864 (2010)

59. Larson, E.C., Chandler, D.M.: Most apparent distortion: full-reference image quality assessment and the role of strategy. J. Electron. Imaging **19**(1):011,006-1–011,006-21 (2010)
60. Li, C., Bovik, A.C.: Content-partitioned structural similarity index for image quality assessment. Signal Process. Image Commun. **25**(7), 517–526 (2010) (special Issue on Image and Video Quality Assessment)
61. Li, Q., Wang, Z.: General-purpose reduced-reference image quality assessment based on perceptually and statistically motivated image representation. In: 15th IEEE International Conference on Image Processing, 2008. ICIP 2008, pp. 1192–1195 (2008)
62. Liu, H., Heynderickx, I.: Studying the added value of visual attention in objective image quality metrics based on eye movement data. In: 16th IEEE International Conference on Image Processing (ICIP), pp. 3097–3100 (2009)
63. Liu, H., Redi, J., Alers, H., Zunino, R., Heynderickx, I.: Efficient neural-network based no-reference approach to an overall quality metric for jpeg and jpeg2000 compressed images. J. Electron. Imaging **20**, 043,007-(1–15) (2011)
64. Lubin, J.: A visual discrimination model for image system design and evaluation. In: Peli, E. (ed.) Visual Models for Target Detection and Recognition, pp. 207–220. World Scientific Publisher, Singapore (1995)
65. Lundstrom, C.: Technical report: Measuring digital image quality. Technical report, Linkping UniversityLinkping University, Visual Information Technology and Applications (VITA), The Institute of Technology (2006)
66. Marziliano, P., Dufaux, F., Winkler, S., Ebrahimi, T.: A no-reference perceptual blur metric. In: IEEE 2002 International Conference on Image Processing, pp. 57–60 (2002)
67. Mittal, A., Moorthy, A.K., Bovik, A.C.: No-reference image quality assessment in the spatial domain. IEEE Trans. Image Process. **21**(12), 4695–4708 (2012)
68. Mittal, A., Soundararajan, R., Bovik, A.C.: Making a completely blind image quality analyzer. IEEE Signal Process. Lett. **20**, 209–212 (2013)
69. Moorthy, A., Bovik, A.: Visual importance pooling for image quality assessment. IEEE J. Sel. Top. Signal Process. **3**(2), 193–201 (2009)
70. Moorthy, A., Bovik, A.: Blind image quality assessment: from natural scene statistics to perceptual quality. IEEE Trans. Image Process. **20**(12), 3350–3364 (2011a)
71. Moorthy, A., Bovik, A.: Visual quality assessment algorithms: what does the future hold? Multimed. Tools Appl. **51**, 675–696 (2011b)
72. Muijs, R., Kirenko, I.: A no-reference blocking artifact measure for adaptive video processing. In: Proceedings of the 13th European Signal Processing Conference 2005 (2005)
73. Ninassi, A., Le Meur, O., Le Callet, P., Barba, D.: Task impact on the visual attention in subjective image quality assessment. In: Proceedings of the 14th European Signal Processing Conference, Eurasip EUSIPCO (2006)
74. Ninassi, A., Le Meur, O., Le Callet, P., Barbba, D.: Does where you gaze on an image affect your perception of quality? Applying visual attention to image quality metric. In: IEEE International Conference on Image Processing, 2007. ICIP 2007, vol. 2, pp. II-169–II-172 (2007)
75. Nishiyama, M., Okabe, T., Sato, I., Sato, Y.: Aesthetic quality classification of photographs based on color harmony. In: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 33–40 (2011)
76. Ong, E., Lin, W., Lu, Z., Yang, X., Yao, S., Pan, F., Jiang, L., Moschetti, F.: A no-reference quality metric for measuring image blur. In: Proceedings of the Seventh International Symposium on Signal Processing and Its Applications, vol. 14, pp. 469–472 (2003)
77. Pan, F., Lin, X., Rahardja, S., Lin, W., Ong, E., Yao, S., Lu, Z., Yang, X.: A locally adaptive algorithm for measuring blocking artifacts in images and videos. Signal Process. Image Commun. **19**(6), 499–506 (2004)
78. Peli, E.: Contrast in complex images. J. Opt. Soc. Am. **7**, 2032–2040 (1990)
79. Pessoa, A., Falcao, A., e Silva, A., Nishihara, R., Lotufo, R.: Video quality assessment using objective parameters based on image segmentation. In: Proceedings of the SBT/IEEE International Telecommunications Symposium, ITS '98, vol. 2, pp. 498–503 (1998)
80. Pinson, M., Wolf, S.: A new standardized method for objectively measuring video quality. IEEE Trans. Broadcast. **50**(3), 312–322 (2004)
81. Ponomarenko, N., Lukin, V., Zelensky, A., Egiazarian, K., Astola, J., Carli, M., Battisti, F.: A database for evaluation of full reference visual quality assessment metrics. Adv. Mod. Radioelectron. **10**, 30–45 (2009)
82. Rank, K., Lendl, M., Unbehauen, R.: Estimation of image noise variance. IEE Proc. Vis. Image Signal Process. **146**(2), 80–84 (1999)
83. Redi, J.A., Liu, H., Zunino, R., Heynderickx, I.: Interactions of visual attention and quality perception. In: IS&T/SPIE Electronic Imaging 2011 and Human Vision and Electronic Imaging XVI, vol. 7865 (2011)
84. Rehman, A., Wang, Z.: Reduced-reference image quality assessment by structural similarity estimation. IEEE Trans. Image Process. **21**(8), 3378–3389 (2012)
85. Reibman, A., Bell, R., Gray, S.: Quality assessment for super-resolution image enhancement. In: 2006 IEEE International Conference on Image Processing, pp. 2017–2020 (2006)
86. de Ridder, H., Endrikhovski, S.: Image quality is fun: reflections on fidelity, usefulness and naturalness. SID Symposium Digest of Technical Papers, vol. 33, pp. 986–989 (2002)
87. Saad, M., Bovik, A., Charrier, C.: A dct statistics-based blind image quality index. IEEE Signal Process. Lett. **17**(6), 583–586 (2010)
88. Saad, M., Bovik, A., Charrier, C.: Blind image quality assessment: a natural scene statistics approach in the dct domain. IEEE Trans. Image Process. **21**(8), 3339–3352 (2012)
89. Safranek, R., Johnston, J.: A perceptually tuned sub-band image coder with image dependent quantization and post-quantization data compression. In: 1989 International Conference on Acoustics, Speech, and Signal Processing, 1989. ICASSP-89, pp. 1945–1948 (1989)
90. Saha, S., Vemuri, R.: An analysis on the effect of image activity on lossy coding performance. In: Proceedings of the 2000 IEEE International Symposium on Circuits and Systems, 2000. ISCAS 2000, Geneva, vol. 3, pp. 295–298 (2000)
91. Sazzad, Z., Kawayoke, Y., Horita, Y.: Mict image quality evaluation database. http://mict.eng.u-toyama.ac.jp/mict/index2.html (2000)
92. Schloss, K., Palmer, S.: Aesthetic response to color combinations: preference, harmony, and similarity. Atten. Percept. Psychophys. **73**, 551–571 (2011)
93. Seshadrinathan, K., Bovik, A.C.: Motion tuned spatio-temporal quality assessment of natural videos. IEEE Trans. Image Process. **19**(2), 335–350 (2010)
94. Sharma, G.: Digital Color Imaging Handbook. CRC Press Inc, Boca Raton (2002)
95. Sharma, G., Bala, R. (eds.): Digital Color Imaging Handbook, vol. 29. CRC, Boca Raton (2003)
96. Sheikh, H., Bovik, A.: Image information and visual quality. IEEE Trans. Image Process. **15**(2), 430–444 (2006)
97. Sheikh, H., Wang, Z., Cormack, L., Bovik, A.: Live image quality assessment database release 2 (2005)

98. Simoncelli, E.P., Olshausen, B.A.: Natural image statistics and neural representation. Annu. Rev. Neurosci. **24**(1), 1193–1216 (2001)
99. Solli, M., Lenz, R.: Color harmony for image indexing. In: 2009 IEEE 12th International Conference on Computer Vision Workshops (ICCV Workshops), pp. 1885–1892 (2009)
100. Soundararajan, R., Bovik, A.: Rred indices: reduced reference entropic differencing for image quality assessment. IEEE Trans. Image Process. **21**(2), 517–526 (2012)
101. Soundararajan, R., Bovik, A.: Video quality assessment by reduced reference spatio-temporal entropic differencing. IEEE Trans. Circuits Syst. Video Technol. **23**(4), 684–694 (2013)
102. Suthaharan, S.: No-reference visually significant blocking artifact metric for natural scene images. Signal Process. **89**(8), 1647–1652 (2009)
103. TASI: Technical advisory service for images. http://www.tasi.ac.uk/advice/creating/quality.html (1979)
104. Teo, P., Heeger, D.: Perceptual image distortion. In: Proceedings of the IEEE International Conference on Image Processing, 1994. ICIP-94, vol. 2, pp. 982–986 (1994a)
105. Teo, P., Heeger, D.J.: Perceptual image distortion. In: Proceedings of the SPIE, pp. 982–986 (1994b)
106. Thurstone, L.L.: A law of comparative judgement. Psychol. Rev. **34**, 273–286 (1927)
107. Tong, Y., Konik, H., Cheikh, F., Tremeau, A.: Full reference image quality assessment based on saliency map analysis. J. Imaging Sci. **54**(3), 30,503-1–30,503-14 (2010)
108. Torgerson, W.: Theory and Methods of Scaling. Wiley, Ney York (1958)
109. Torralba, A., Oliva, A.: Statistics of natural image categories. Netw. Comput. Neural Syst. **14**(3), 391–412 (2003)
110. Tourancheau, S., Autrusseau, F., Sazzad, Z., Horita, Y.: Impact of subjective dataset on the performance of image quality metrics. In: 15th IEEE International Conference on Image Processing, 2008. ICIP 2008, pp. 365–368 (2008)
111. Vlachos, T.: Detection of blocking artifacts in compressed video. Electron. Lett. **36**(13), 1106–1108 (2000)
112. VQEG: Vqeg final report of fr-tv phase ii validation test. Technical report, Video Quality Experts Group (VQEG) (2003)
113. Wang, Z., Bovik, A.: Mean squared error: love it or leave it? A new look at signal fidelity measures. IEEE Signal Process. Mag. **26**(1), 98–117 (2009)
114. Wang, Z., Li, Q.: Information content weighting for perceptual image quality assessment. IEEE Trans. Image Process. **20**(5), 1185–1198 (2011)
115. Wang, Z., Simoncelli, E.P.: Reduced-reference image quality assessment using a wavelet-domain natural image statistic model. In: Proceedings of the SPIE Human Vision and Electronic Imaging, vol. 5666, pp. 149–159 (2005)
116. Wang, Z., Bovik, A., Evans, B.: Blind measurement of blocking artifacts in images. In: Proceedings of the IEEE International Conference on Image Processing, pp. 981–984 (2000)
117. Wang, Z., Sheikh, H., Bovik, A.: No-reference perceptual quality assessment of jpeg compressed images. In: Proceedings of the 2002 International Conference on Image Processing, vol. 1, pp. I-477–I-480 (2002)
118. Wang, Z., Simoncelli, E., Bovik, A.: Multiscale structural similarity for image quality assessment. In: Conference Record of the Thirty-Seventh Asilomar Conference on Signals, Systems and Computers, 2004, vol. 2, pp. 1398–1402 (2003)
119. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004a)
120. Wang, Z., Simoncelli, E.P., Hughes, H.: Local phase coherence and the perception of blur. In: Advances in Neural Information Processing Systems, NIPS03, pp. 786–792. MIT Press, Cambridge (2004b)
121. Watson, A.B.: DCT quantization matrices visually optimized for individual images. In: Allebach, J.P., Rogowitz, B.E. (eds.) Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, vol. 1913, pp. 202–216 (1993)
122. Watson, A.B., Borthwick, R., Taylor, M.: Image quality and entropy masking. In: SPIE Human Vision and Electronic Imaging Conference, vol. 3016, pp. 2–12 (1997)
123. Wayne, S.: Quality control circle and company wide quality control. Qual. Prog. **16**(10), 14–17 (1983)
124. Wee, C.Y., Paramesran, R., Mukundan, R., Jiang, X.: Image quality assessment by discrete orthogonal moments. Pattern Recognit. **43**(12), 4055–4068 (2010)
125. Winkler, S., Süsstrunk, S.: Visibility of noise in natural images. In: Proceedings of the IS&T/SPIE Electronic Imaging 2004: Human Vision and Electronic Imaging IX, vol. 5292, pp. 121–129 (2004)
126. Wu, H., Yuen, M.: A generalized block-edge impairment metric for video coding. IEEE Signal Process. Lett. **4**, 317–320 (1997)
127. X-Rite: X-Rite ColorChecker Classic. http://xritephoto.com/ph_product_overview.aspx?ID=1192. Accessed 09 Feb 2012
128. Ye, P., Doermann, D.: No-reference image quality assessment using visual codebooks. IEEE Trans. Image Process. **21**(7), 3129–3138 (2012)
129. Yeganeh, H., Wang, Z.: Objective quality assessment of tone-mapped images. IEEE Trans. Image Process. **22**(2), 657–667 (2013)
130. Yeganeh, H., Rostami, M., Wang, Z.: Objective quality assessment for image super-resolution: a natural scene statistics approach. In: 2012 19th IEEE International Conference on Image Processing (ICIP), pp. 1481–1484 (2012)
131. Yen, R., Zektser, G.: A new approach for measuring facial image quality. Def. Stand. Progr. J. 13–25 (2008)
132. Yendrikhovskij, S.: Image quality: between science and fiction. In: PICS, pp. 173–178 (1999)
133. Zhang, D.Q., Chang, S.F.: Detecting image near-duplicate by stochastic attributed relational graph matching with learning. In: Proceedings of the 12th Annual ACM International Conference on Multimedia, MULTIMEDIA '04, pp. 877–884 (2004)
134. Zhang, L., Zhang, D., Mou, X., Zhang, D.: Fsim: a feature similarity index for image quality assessment. IEEE Trans. Image Process. **20**(8), 2378–2386 (2011)
135. Zhang, X., Wandell, B.A.: A spatial extension of cielab for digital color-image reproduction. J. Soc. Inf. Disp. **5**(1), 61–63 (1997)
136. Zhu, X., Milanfar, P.: Automatic parameter selection for denoising algorithms using a no-reference measure of image content. IEEE Trans. Image Process. **19**(12), 3116–3132 (2010)