UNIVERSITÀ DEGLI STUDI DI MILANO-BICOCCA

Facoltà di Scienze Matematiche, Fisiche e Naturali

Dipartimento di Informatica, Sistemistica e Comunicazione

Dottorato di Ricerca in Informatica - XXIII Ciclo

Content Based No-Reference Image Quality Metrics

Ph.D. Dissertation of: Fabrizio Marini

Supervisor: Prof. Raimondo Schettini

Tutor: Prof. Fabio Stella

Ph.D. Coordinator: Prof.ssa Stefania Bandini

Anno Accademico 2010-2011

Contents

Abstract

Images are playing 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 anytime. Since digital images are subject to a wide variety of distortions during acquisition, processing, compression, storage, transmission and reproduction; it becomes necessary to assess the Image Quality. In this thesis, starting from an organized overview of available Image Quality Assessment methods (Chapter 1), some original contributions in the framework of No-reference image quality metrics are described.

No-Reference metrics are also called blind as they assume that image quality can be determined without a direct comparison between the original and the processed images. To this end, No-Reference metrics are designed to identify and quantify the presence of specific processing distortions that may exist in the evaluated image. To estimate the presence of a defect or artifact, the properties of the artifact as well as the effects that it produces on the low level components of the image (edges, homogeneous areas, etc) should be modeled and characterized. For each image artifact, several metrics and benchmark databases are often available in the literature. There are, however, some artifacts that have not been studied yet. One of these is due to demosaicing and it is investigated in Chapter 2. The demosaicing operation consists of a combination of color interpolation and anti-aliasing algorithms and converts a raw image acquired with a single sensor array, overlaid with a color filter array, into a full-color image. The most prominent artifact generated by demosaicing algorithms is called zipper. The zipper artifact is characterized by segments (zips) with an On-Off pattern. In this chapter I describe psycho-visual experiments to analyze the perceived distortions produced by different demosaicing algorithms. To this end, I have generated a proper dataset with nine different degrees of distortions, using three color

interpolation algorithms combined with two anti-aliasing algorithms. In this thesis I propose a new metric based on measures of blurriness, chromatic and achromatic distortions that fit the psycho-visual data collected during the subjective experiments.

Typically, No-Reference metrics are designed to measure specific artifacts using a distortion model. Some psycho-visual experiments have shown that the perception of distortions is influenced by the amount of details in the image's content, suggesting the need for a content weighting factor. This dependency is coherent with known masking effects of the human visual system. In Chapter 3, I focus on the blocking distortion of JPEG compressed images and show that information about the visual content of the image (encoded as wavelets descriptors) can be exploited to improve the estimation of the quality of JPEG compressed images.

In Chapter 4 I focus on No-Reference metrics for sharpness. In the methods available in the literature, after detecting the edge pixels, the sharpness measure is defined for each edge pixel. I have observed that in some cases this global measure is not representative of the real sharpness of the images. This fact is mainly due to the image noise that interferes with the measure at pixel-level. Performing the measure on a set of edge pixels can mitigate this problem. In this chapter, I propose a method that automatically selects edge segments, and permits to evaluate image sharpness on more reliable data. Moreover a novel sharpness metric for natural images, inspired by the slanted edge technique used in case of synthetic images is introduced. This metric makes it possible to cope with noise influence providing more reliable estimations.

In Chapter 5 I present a modular No-Reference Image Quality tool. This tool addresses natural images in general where signal content and distortion may not be clearly separated. For this scope, the tool gives a structured view of a large collection of objective metrics that are available for the different distortions within an integrated framework. The tool permits to apply the metrics not only globally but also locally to different regions of interest of the image. I have observed that a criterion to define the "best metric" for each distortion does not exist. This ideal metric should therefore take into account the pictorial content of the image. As this could be a difficult task, our tool allows the operator to use his prior knowledge to select a region where the signal content does not affect the measure of the distortion. In the conclusions the main contributions of this thesis are outlined the future work discussed.

Chapter 1

Survey on Image Quality Assessment

Quality, in general, has been defined as the "totality of characteristics of a product that bear on its ability to satisfy stated or implied needs" [48]; "fitness for (intended) use" [54]; "conformance to requirement" [27]; "user satisfaction" [130]. These definitions and their numerous variants could fit digital IQ as suggested 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" [109]. According to the International Imaging Industry Association [107], IQ is the perceptually weighted combination of all visually significant attributes of an image when considered in its marketplace or application. We must, in fact, consider the application domain and expected use of the image data. An image, for example, could be used just as a visual reference to an item in the digital archive; and although IQ has not been precisely defined, we can reasonably assume that in this case IQ requirements are low. On the contrary if the image were to "replace" the original, IQ requirements would be high.

Taking into account that images are not necessarily processed by a human observer, we can consider the quality of an image as the degree of adequacy to its function/goal within a specific application field. Given a specific domain and task, there are several factors that may influence the results and therefore the perceived IQ. These are: scene geometry and lighting conditions, imaging device (HW and embedded SW), image processing and transmission, rendering device (HW and embedded SW), the observer's adaptation

state and viewing conditions, the observers' previous experiences, preferences and expectations.

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. Since an image is the result of the optical imaging process, which maps physical scene properties onto a two-dimensional luminance distribution, it encodes important and useful information about the geometry of the scene and the properties of the objects located within this scene [50, 113, 134].

Janssen and Blommaert [51] regard the visuo-cognitive processing not as an isolated process but instead as an essential stage in human interaction with the environment. According to these authors, the quality of an image is the adequacy of this image as input to visual perception and this adequacy is given by the discriminability and identifiability of the items depicted in the image. In Figure 1.1 their schematic overview of the interaction process is shown. Images are the carriers of information about the environment, and serve as input to visual perception. The result of visual processing is used as input to cognition (for tasks requiring interpretation of scene content) or as input to action (for example in navigation, where the link between perception and action is mostly direct). Since action will in general result in a changed status of the environment, the nature of the interaction process is cyclic.

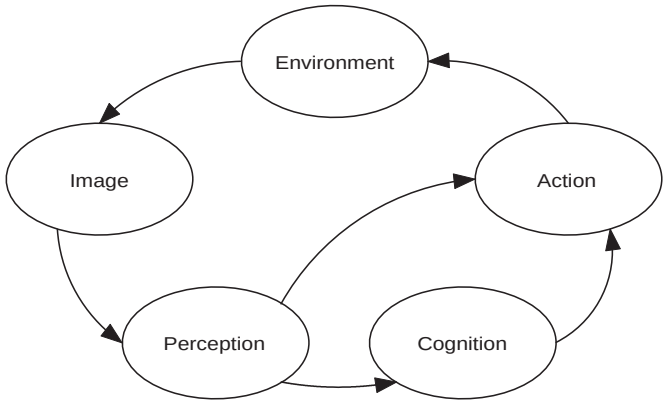


Figure 1.1: Schematic overview of the interaction process by Janssen and Blommaert 2000.

As above mentioned, the concept of IQ does not always correspond to image fidelity. The FUN IQ model [29] depicted in Figure 1.2 assumes the

existence of three major dimensions to determine IQ:

- Fidelity is the degree of apparent match of the acquired/reproduced images 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 require a high fidelity of the images with respect to the originals. Genuineness and faithfulness are sometimes used as synonymous of Fidelity [107]. Dozens of books and thousands of papers have been written about image fidelity and image reproduction e.g. [98].

- Usefulness is the degree of apparent suitability of the acquired/reproduced image with respect to a specific task. In many application domains, such as medical or astronomical imaging, image processing procedures can be applied to increase the image usefulness [40]. These processing steps have an obvious impact on Fidelity.

- Naturalness is the degree of apparent match of the acquired/reproduced images 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 those 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.

It should be noted that, in general, the quality dimensions are not independent however the overall IQ is usually evaluated as a single number weighting the individual components. These weights depend on the specific image data type and on its function/goal within a specific application field.

The goal of the present article is not only to review the state of the art of the different IQA methods (see for example [8]) but also to guide a non-expert user in the choice and/or design of a workflow chain that has to make use of IQ metrics to validate the corresponding output products.

In what follows we review the literature on objective IQ methods and we analyze how and when these different kind of metrics can be applied to a generic image workflow chain. In section 1.1 these metrics are classified and briefly described. In section 1.2 we consider region-of-interest based IQA that has become nowadays an active topic of research. In section 1.3 the available

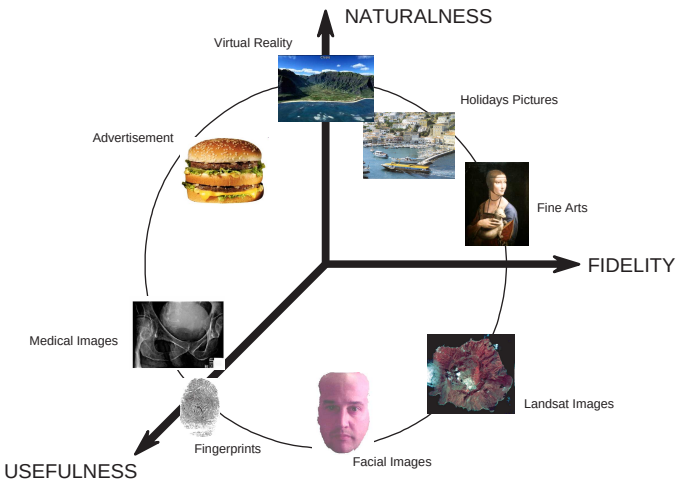


Figure 1.2: Fidelity-Usefulness-Naturalness (FUN) image quality representation (Ridder and Endrikhovski, 2002.

databases for IQA are summarized. In sections 1.4 and 1.5 a generic image workflow chain, starting from a real scene to be captured, is shown and the different IQ metrics that could eventually be applied are analyzed. Finally, the conclusions are drawn in section 6.

1.1 Image Quality Assessment Approaches

Different criteria can be used in order to classify the IQA approaches. The first criteria is to divide the methods into two big groups: direct versus indirect methods. Direct approaches take into account the quality of the image itself while indirect approaches, consider the quality of the image with respect to the performances reached by the application that uses the images. Within each of these macro groups, we can categorize the approaches in subjective versus objective methods.

1.1.1 Direct Image Quality Approaches

Subjective methods are based on psychological experiments involving human observers. Different techniques can be used like single or double stimulus and pair wise comparison among others. Standard psychophysical scaling tools

for measure subjective IQ are now available and described in some Standards, such as ITU-R BT.500-11 ([49, 31, 120]). The involvement of real people who view the images for assessing their quality requires that all the factors that influence perception should be taken into account and strict protocols have to be adopted. Notwithstanding effectiveness of subjective approaches, their efficiency is very low compared to objective ones.

Moreover, subjective quality issues could be discarded if the image usage does not require the user involvement, or if the observer could be substituted by a computational model. This has led the research towards the study of objective IQ measures not requiring human interaction.

Objective methods compute suitable metrics directly from the digital image. Within this group, the methods can be classified according to many different criteria. According to the availability of the original image the methods can be Full Reference (FR), No Reference (NR) and Reduced Reference (RR) [8]. Taking into account the phylosophy followed when constructing the algorithm, the methods can be classified as bottom-up or top-down. If we consider the application scope, they can be general purpose or context-dependent. Engeldrum [32] presented a general taxonomy classifying the models as detection/recognition versus beauty context and physically vs. ness-based. In what follows we make a summary of well known methods belonging to the different categories FR, NR and RR.

- Full-reference (FR) metrics (see Figure 1.3) perform a direct comparison between the image under test and a reference or "original" in a properly defined image space.

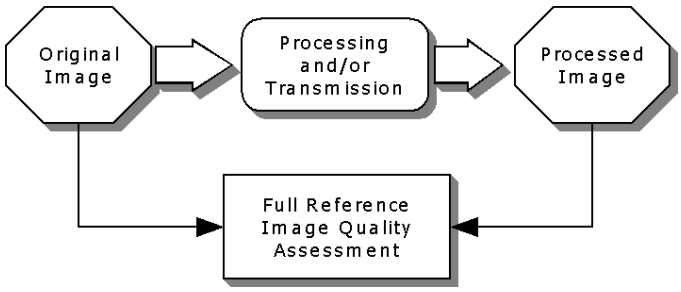


Figure 1.3: Image quality assessment approaches: Full Reference.

Having access to an original is a requirement of the usability of such metrics. Among the quality dimensions previously introduced, only

image fidelity can be assessed. The simplest FR metric is the Mean Square Error (MSE) or Peak to Signal Noise Ratio (PSNR). Even if it is the most used, in general does not correlate with subjective results. Error sensitivity frameworks follow a strategy of modifying the MSE measure so that errors are penalised in accordance with their visibility. For example in Figure 1.4 is shown an example of how the perceived image quality is strongly influenced by the distortion visibility. To the original image (Figure 1.4a) is applied a Gaussian noise to the sky/clouds (Figure 1.4b) and to the sand/rocks (Figure 1.4c) region of the image. When the distortion is applied to the sand/rock region, it is less noticeable. 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.

Figure 1.4: Example of how the perceptual quality is influenced by the visibility of the distortion. a)The original image. The same distortion (Gaussian noise) is applied to the top (b) and bottom (c) regions of the image. The image in (c) is perceived to have higher quality than the image in (b).

The evaluation of the visibility is accomplished by modeling some aspects of the Human Visual System (HVS) like the channel decomposition, Contrast Sensitivity and Point Spread functions among others [28, 110, 68, 93]. Zhang and Wandell [136] proposed an error-visibility metric, s-CIElab, which is an extension of the CIE color difference equations to complex stimuli by means of some spatial filters utilized to model the Contrast sensitivity function. All these techniques are bottom-up like approaches. Recently, Laparra et al. [60] extended the divisive normalization metric originally proposed by Teo and Heeger [110] and that is based on the standard psychophysical and physiological model that describes the early visual processing.

On the other hand, the Structural Similarity Measure (SSIM) [123]

uses a different concept of IQ. Starting from the assumption that natural image signals are highly structured, a measurement of structural dissimilarity (or distortion) should provide a good approximation to perceived IQ. This is a top-down approach as it assumes that finding the structure is the goal for the cognitive process. The structural information in an image is define as those attributes that represent the structure of objects in the scene, independently of the average luminance and contrast. Their system separates the task of similarity measurement into three comparisons: local luminance, local contrast and local structure.

The IQ has also been addressed within the information-theoretic approach. Sheikh and Bovik [102] proposed the Visual Information Fidelity Index (VIF). This model uses natural scene statistics and the index quantifies loss of information due to the distortions present in the image. A brief summary of these FR methods is presented in table 1.1.

- No-reference (NR) metrics (see Figure 1.5) are also called blind metrics and assume that IQ can be determined without a direct comparison between the original and the processed images.

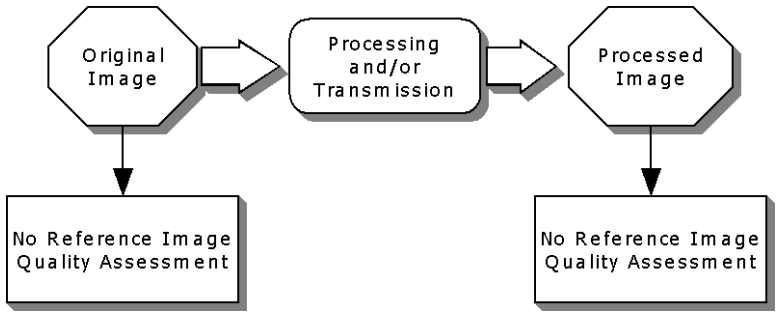


Figure 1.5: Image quality assessment approaches: No-Reference.

Theoretically, it is possible to measure the quality of any visual contents. In practice, some information about the application domain, requirements and users' preferences are required to contextualize the quality measures. NR metrics are designed to identify and quantify the presence of specific processing distortions that may exist in the evaluated image. To estimate the presence of a defect or artifact produced

by some imaging processing on the image, we need to characterize the properties of the artifact as well as the effects that it produces on the low level components of the image (edges, homogeneous areas, etc). Different types of artifacts can be considered like blurriness, graininess, blocking, lack of contrast and lack of saturation or colorfulness among others [43]. Blind methods can be classified as application-dependent since they are defined to handle with one or few specific artifact types.

Some of the blind methods are carried out in the frequency domain (like [22] for example) and make use of the common statistical characteristics of the power spectra of natural images [114] in order 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 [105].A brief summary of different NR methods is presented in table 1.2.

- Reduced-reference (RR) metrics (see Figure 1.6) lie between FR and NR metrics.

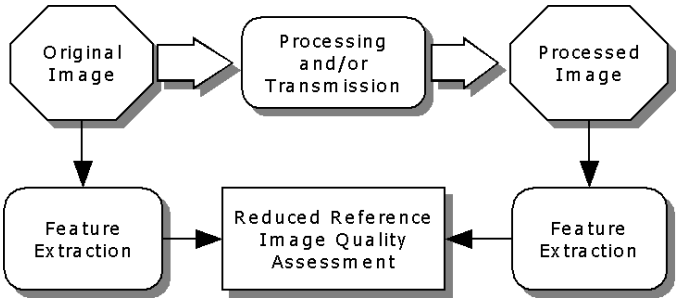
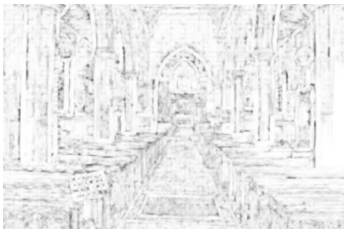


Figure 1.6: Image quality assessment approaches: Reduced Reference.

They are designed to predict perceptual IQ with only partial information about the reference images. The methods extract a number of features from both the reference and the image under test. These features are used as a surrogate of all the information of images and image comparison is based only on the correspondence of these features [119, 131, 126, 58, 94, 64, 14]. Therefore, only image fidelity can be assessed. RR metrics are useful to track the degree of visual degradation of video data that are transmitted through communication networks. Compared with FR and NR, few RR methods are available





$MSE = 28.5721$
$SSIM = 0.9204$

$\Delta E_{94}(c1)$=3.47
$\Delta E_{94}(c2)$=2.82
$\Delta E_{94}(c3)$=1.04
...

Figure 1.7: An example of quality assessment outputs. Top row: the original image (left); a JPEG-compressed version of the original image (right). Bottom row: the SSIM error map (left) where darker values indicate higher errors; two quality indexes (center); errors in the reproduction of some reference colors (right). In the example a color chart is acquired with the original subject and used as color fidelity reference.

in the literature. For the reduced description of the image, the methods in general use features describing the image content or distortion-based features. These features must be then coded and transmitted with the compressed image data produced by the coder using a side channel with low transmission error. In table 1.3 a brief summary of RR methods is presented.

The output of the assessment procedure can be a number, a set of numbers or an image error map (Figure 1.7). The map can then be used to precisely locate where the imaging processing procedures degrade the image.

To validate the objective methods, the metrics are evaluated on a database and subjective tests are carried out simultaneously using standard psychophysical scaling tools. Both objective and subjective results are then compared through different performance metrics such as correlation coefficients and Spearman rank order correlation coefficient [120]. It should be noted that, since the subjective quality score is a single numerical value also the objective quality measure must be expressed using a single value.

1.1.2 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 pixel or feature values. Image quality can also be indirectly assessed by:

- Quantifying the performance of an image-based task. This can be done manually by domain experts and/or automatically by a computational system. For example, in a biometrics system an image of a face is of good quality if the person can be reliably recognized. This can be done by manually inspecting each image acquired and evaluating if the pose satisfies the application constraints (e.g. non-occluded face) or enforced by law requirements (e.g. open eyes). Image distortions that are irrelevant for the task can therefore go unnoticed or simply ignored by the observer. We can consider these as irrelevant. The quality evaluation could be done by a face recognition algorithm that automatically processes each images and assesses the fulfillment of the constraints and requirements [69].

- 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 and ad-hoc designed software tools (e.g. [45]), but these measures alone are not sufficient to fully assess IQ. The Camera Phone Image Quality (CPIQ) Initiative of the International Imaging Industry Association (I3A) suggests both objective and subjective characterizations procedures [107].

1.2 Visual Perception and Quality Assessment

The HVS is specialized and tuned to recognize the features that are most important for human evolution and survival. On the other hand, there are other image features that humans cannot distinguish or that are easily overlooked. There are some intrinsic limitations of the HVS relevant to IQA like luminance sensitivity, contrast sensitivity, and texture masking [129, 81]. These

limitations make quality assessment highly dependent on the image contents. Moreover, subjective experiences and preferences may influence the human assessment of IQ; for example, it has been shown that the perceived distortions are dependent on how familiar the test person is with the observed image. IQA is also affected by the user's task, (see for e.g [69] and [30]): passive observation can be reasonably assumed when the observer views a vacation image, but not x-rays for medical diagnosis.

The cognitive understanding and interactive visual processing, like eye movements, influence the perceived quality of images in a top-down way [123]. If the observer is provided with different instructions when evaluating a given image, he will give different scores to the same image depending on those instructions. Prior information regarding the image contents or fixation, may therefore affect the evaluation of the IQ. It is also well known that one of the objects attracting most of our attention are people and especially human faces. If there are faces of people in a scene, we will look at them immediately and, because of our familiarity with peoples faces, we are very sensitive to distortions or artifacts occurring in them. In eye-tracking experiments, it has been found that the eye positions recorded under the free-task condition are different from the regions recorded when people fixate while judging IQ [121]. For example it has been observed that in the case of blurring and white noise, the fixations while rating IQ do not change with respect to the task-free condition but in the case of compression artifacts, these can influence fixations, depending on the amount of distortion. Therefore, visual attention and gaze direction appear as two important factors that may influence human perception of IQ.

Besides the error sensitivity frameworks mentioned in section 1.1 that model some aspects of the HVS ([28, 110, 68, 93]), region-of-interest based IQA has become nowadays an active topic of research. In order to investigate if artifacts are likely more annoying in salient regions than in other areas, many different IQA experiments where done during which eye movements were also recorded and the corresponding databases have been generated (see following section). Up to date, controversial results exist for including saliency maps in FR methods. The basic idea is to assign visual importance weights to the MSE, PSNR, SSIM or VIF metrics, giving more importance to the degradation appearing on the salient areas. Some authors ([61, 112, 70, 71, 74]) showed that better agreement with subjective scores can be produced for IQA metrics when saliency maps are taken into account in the metrics' evaluation. In a similar way, image regions can be weighted in the IQ FR

metrics according to some visual properties like edge regions vs. texture regions vs. plain regions as in [63, 115].

On the other hand, others claim that for example, MSE and SSIM do not show a clear improvement [76, 135]. In particular, the results from Ninassi et al. suggest that the way to take into account the visual attention cannot be limited to a simple spatial pooling. Another reason might be that the viewers had enough time to look at all parts of the image when evaluating its quality, such that the influence of attention regions on the overall quality of whole image would not be great.

With respect to the integration of saliency maps and RR or NR methods, few research has been done up to date. A no-reference perceptual sharpness quality metric was proposed by Sadaka et al. [92] that integrates saliency maps with a blur distortion metric, giving more weight to salient edges and penalizing those edges appearing in non-attended regions. Their simulation results showed an increased correlation with the MOS of the subjective measures.

Therefore, taking into account the cognitive behavior presents a challenge to the quality assessment comunity that will certainly continue to be a focus of research for the next years.

1.3 Dataset for Image Quality Estimation

In order to validate the different algorithms' results with human subjective judgments of quality, different database are available to test the algorithms' performance. Among the most frequently used we can cite: LIVE [104], MICT [95], IVC [13] and TID2008 [84]. Even though it is a rather small dataset, the A57 [15] database is also available. The Visual Attention for Image Quality database (VAIQ) [33] facilitates the incorporation of visual attention models into IQ metrics that are designed based on the IVC, LIVE, and MICT databases.

There exist also other kind of databases like the Database Of Visual Eye movementS (DOVES) [116], that is a collection of eye movements from human observers as they view natural calibrated images. Using the DOVES database, [86] evaluated the contributions of four foveated low-level image features (luminance, contrast and bandpass outputs of both luminance and contrast) in drawing fixations of observers. They discovered that image patches around human fixations had, on average, higher values of each of

these features than image patches selected at random. Using these measurements, they developed an algorithm that selects image regions as likely candidates for fixation called GAFFE for Gaze-Attentive Fixation Finding Engine [86].

Where eye tracking devices are not available, models of saliency can be used to predict fixation locations. Most saliency approaches are based on bottom-up computation that do not consider top-down image semantics and often do not match actual eye movements. To address this problem, Judd et al. (2009) collected eye tracking data and used this database as training and testing examples to learn a model of saliency based on low, middle and high-level image features.

In table 1.4 we present a brief description of the above cited databases.

1.4 Image Production Workflow

In Figure 1.8 a generic image workflow chain is shown. It starts with a real scene to be captured by a digital image. The scene is acquired by a proper device (e.g. digital camera or scanner) that performs all the processing steps aimed to produce a digital representation of the scene. Examples of these processing steps are geometric transformation, gamma correction, color adjustments, etc. Imaging metadata can be automatically embedded in the image header (e.g. EXIF) by the imaging device and may include some information such as maker and model of the camera, device settings and preprocessing, date and time, Time Zone offset, and GPS Information. Other metadata are usually added both for catalogue and retrieval purposes. These metadata can include both textual annotations inserted by cataloguers in the context of the application or automatically computed image representations (numerical or alphanumerical features) of some attributes of the digital images. These features are usually related to visual characteristics or be related to symbolic, semantic, or emotional image interpretation, and can be used to derive other information about the image contents [21, 19]. The metadata schema is usually set at the beginning of the digitalization stage and is based on application needs and the workflow requirements. Once the image is acquired, a validation procedure can be applied. This procedure is aimed to have an initial assessment of the suitability and/or quality of the image with respect to the application needs. For example, a manual inspection can be performed in order to check if the whole scene has been correctly acquired

or satisfies some constraints. In some cases, the validation step can be automatically performed using suitable algorithms borrowed from the pattern recognition field.

Images passing through the validation step may have extra ancillary information added to them (e.g. identity of a subject). If required, the image can be further processed in order to increase its usefulness for the task at hand (e.g. contrast enhancement or binarization) or in order to allow more efficient transmission and storage. Again, extra information can be added. The image thus obtained can be finally rendered taking into account both the user's device characteristics and the viewing conditions. These characteristics will not be considered if the images are automatically processed by a computational system. Every element in the workflow chain affects the quality of the resulting images. IQ can be assessed in the different processing stages using one of the approaches discussed in the previous section.

In the IQ literature little attention is given to the scene contents. The scene is composed of the contents itself (a face, for example), and the viewing/acquisition environment: geometry, lighting and surrounding. We call *scene gap* the lack of coincidence between the acquired and the desired scene. The scene gap should be quantified either at the end of the acquisition stage or during the validation stage (if any). The scene gap can be considered recoverable if subsequent processing steps can correct or limit the information loss or corruption in the acquired scene. It is unrecoverable if no suitable procedure exists to recover or restore it. The recoverability of the scene gap is affected by the image domain. When narrow image domains are considered (e.g. medical X-Rays images), to have limited and predictable variability of the relevant aspects of image appearance, it is easier to devise procedures aimed to automatically detect or reduce the scene gap. When broad image domains are considered, it is very difficult and in many cases impossible to automatically detect, quantify and recover the scene gap.

The characteristics of the imaging devices have an obvious impact on the quality of the acquired images. The hardware (sensors and optics) and software components (processing algorithms) of the device may be very articulated and complex. Their roles can be to keep image fidelity as much as possible, improve image usefulness, naturalness, or suitable combinations of these quality dimensions. We call *device gap* the lack of coincidence between the acquired image and the image as acquired by an ideal device properly defined, or chosen and used. The characteristics of the devices to be used must be carefully evaluated in order to make the best cost-performance choice in

accordance to what it is needed for the application at hand and to how the image must be accessed, processed and used. Figure 1.9 shows the generic image workflow chain with the indication of where the different IQA approaches are applied. The FR quality assessment metrics can be applied only when two digital images are available.

Table 1.1: Full Reference Methods

FR Method	What do they measure?	Brief description
MSE and PSNR	closeness to original	Does not take into account HVS characteristics. It is the simplest and oldest measure. No parameters are needed.
Error sensitivity framework. Different models: Daly [28], Lubin [68], Safranek and Johnston [93], Teo and Heeger [110], Watson [128]	Closeness to the original.	Bottom-up approach: simulate functional properties of the HVS. Consist essentially of four modulues: 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.
Structure Similarity Index (SSIM) [123]	Closeness to the original	Top-down approach: 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.
Visual Information Fidelity Index (VIF) [102]	Information shared between the two images	Information fidelitybased approach. The construction of the VIF Index relies on modeling of the statistical image source, the image distortion channel, and the human visual distortion channel. Different parameters have to be estimated.
Spatial-CIELAB [136]	Color differences	Extention of the CIELAB color metric. The image data is transformed into an opponent color space, followed by a CSF spatial filtering. An error map is evaluated. Different parameters have to be estimated.
Discrete orthogonal moments [132]	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.
Divisive normalization metric [60]	Closeness to the original	Based on divisive normalization models in Discrete Cosine Transform and wavelet domains. The general idea to assess the perceptual distance between two images is to compute the q-norm Euclidean distance in the image representation at the primary visual cortex, as suggested in [110].

Table 1.2: No Reference Methods

NR Method	Artifacts	Brief description
Marziliano et al. [35]	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.
Wang et al. [4]	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.
Wang et al. [125]	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.
Bovik and Liu [9]	Blockiness	Discrete Cosine Transform-domain algorithm. Blocking artifact modeled as a 2-D step function. Luminance and texture masking taken into account.
Pan et al. [79]	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.
Vlachos [117]	Blockiness	Designed in the frequency domain. The blockiness measure is defined as the ratio between intra- and inter-block similarity.
Suthaharan [108]	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. Following Wainwright et al. (2002), the metric quantifies visual impairment by altering the spatial frequencies of the channels in order to standardize its sensitivity output such that it is independent from other channels.
Hasler and Susstrunk [43]	Colorfulness	Study of the distribution of the image pixels in the CIELab colour space, assuming that the colourfulness can be represented by a linear combination of a subset of different quantities (standard deviation and mean of saturation and/or chroma). Parameters found by maximising the correlation between experimental data and the metric.
Peli [82]	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).
Wang and Simoncelli [127]	Blur	Defined in the frequency domain. Blur is interpreted as a disruption of local phase. They show that precisely localized features such as step edges result in strong local phase coherence structures across scale and space in the complex wavelet transform domain, and blurring causes loss of such phase coherence. 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.
Ong et al. [78]	Blur	The average edge spread in the image is measured by the average extent of the slopes spread of an edge in both the gradients direction and also the direction opposing the gradients direction.

NR Method	Artifacts	Brief description
Ciancio et al. [18]	Blur	An overcomplete wavelet transform of the image is computed. Coefficients of subbands with the same orientation are expected to be located in similar positions. Following Wang and Simoncelli [127], blur will introduce phase incoherence, causing these positions to change from subband to subband. Coefficients are classified as coherent or incoherent based on an adaptive threshold. The blur estimation is calculated as the mean of the standard deviations of the image components associated to the incoherent coefficients.
Choi et al. [16]	Blur and noise	Blur is estimated by the difference between the intensity of current pixel and average of neighbor pixels, the difference is normalized by the average.
Brandao and Queluz [11]	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 Watson [128].
Gabarda and Cristobal [38]	Blur and noise	The method is based on measuring the variance of the expected entropy of a given image upon a set of predefined directions. Entropy can be calculated on a local basis by 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. Noise-free natural images have shown a maximum of this metric in comparison with other degraded, blurred, or noisy versions.
Cohen and Yitzhaky [22]	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.
Winkler and Susstrunk [133]	Noise	Investigate the visibility of noise itself as a target and use 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).
Rank et al. [87]	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.
Corner et al. [25]	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 estimation.
Immerkaer [47]	Noise	Estimates sigma of the normally distributed noise.
Gasparini et al. [39]	Zipper	Demosaicing metric

Table 1.3: Reduced Reference Methods

RR Method	Features	Brief description
Wang and Simoncelli [126]	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 subband 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.
Kusuma and Zepernik [58]	Features describing blocking and blurring artifacts.	Hybrid IQ metric. The importance of blocking effect is computed using Wang and Bovik method [4] and the importance of blurring is measured using Marziliano's method [35].
Saha and Vemuri [94]	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 subbands coding schemes.
Li and Wang [64]	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.
Carnec et al. [14]	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 organisational model of the HVS, including important stages of vision (perceptual color space, CSF, psychophysical subband 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.

Table 1.4: Image Quality Databases

Database	Brief description
LIVE [104]	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 [95]	14 reference images, 168 test images, 16 observers/image. Distortion types: JPEG and JPEG2000
IVC [13]	10 reference images, 235 test images, 15 observers/image. Distortion types: JPEG, JPEG2000 ; LAR coding ; Blurring
TID2008 [84]	25 reference images, 1700 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.
A57 [15]	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 subbands of a 5-level DWT of the image.
VAIQ [33]	Eye tracking experiments: 42 images, 15 participants. Recorded data per person and image: about 480-540 samples
DOVES [86]	Eye tracking experiments: 101 natural images, 29 participants. The database consists of around 30,000 fixation points
Judd et al. [53]	Eye tracking experiments: 1003 images, 15 viewers.

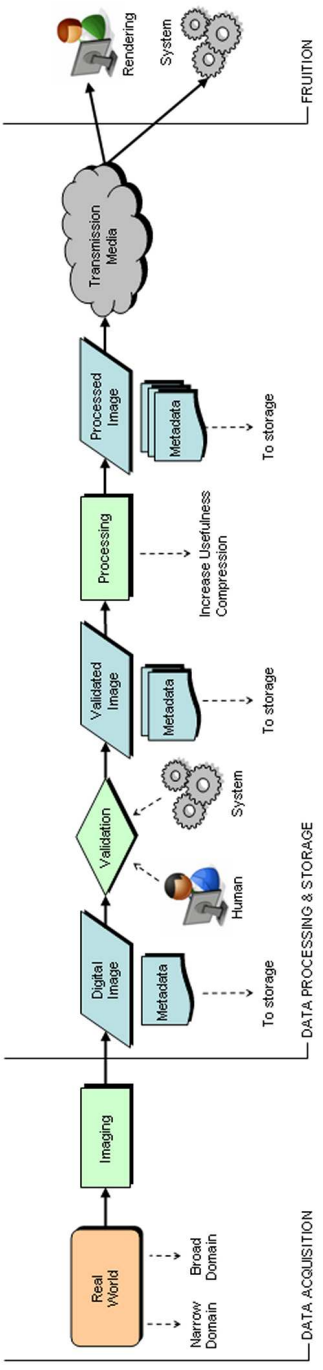


Figure 1.8: A generic image production workflow chain.

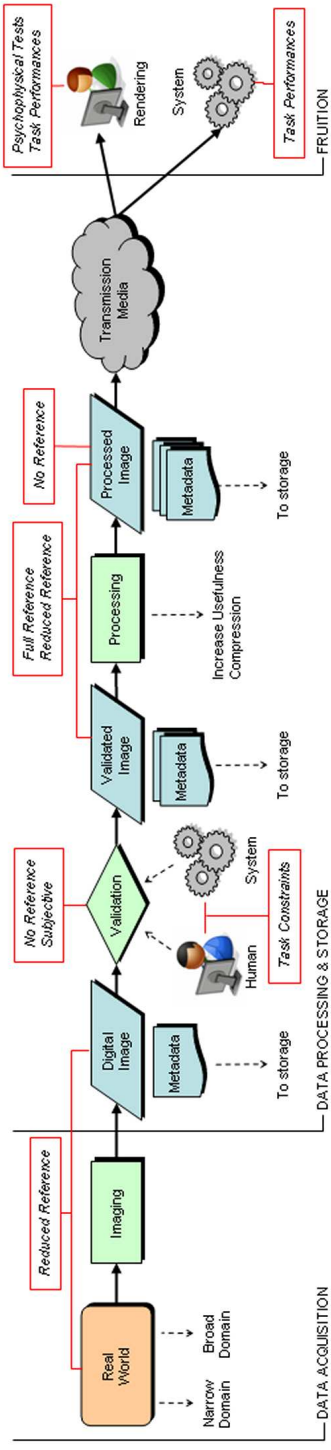


Figure 1.9: Relationship between the image production workflow chain and the image quality assessment approaches.

1.5 Image Reproduction Workflow

For the rendering devices it is important to evaluate the artifacts that their processing pipelines eventually introduce. We call *rendering gap* the lack of coincidence between the actually rendered image and the image as rendered by an ideal (perfect) device properly defined, or chosen and used, for the application at hand. The viewing conditions have a significant influence on the appearance of a rendered image because they can amplify or diminish the visibility of artifacts. This is why all the standards for subjective IQA pay particular attention to this issue. Finally, the observers' previous experiences, preferences and expectations clearly vary and are nearly impossible to standardize. We call *observer gap* the lack of coincidence between the actual observer and the observer the image creator had in mind for a given scope and application. Proper screening and selection of the panel of the observers to be used in the IQA is thus required.

As an example, let us consider the following task: given an input image we would like to predict the overall IQ of the final output print document. The IQA of the image has to be evaluated before printing the document, so that the final product reaches the desired quality level. A generic image reproduction workflow is shown in Figure 1.10. In this reproduction scenario, we assume that the conditions of the image acquisition are unknown so with respect to the scenario in Figure 1.8 no reduce reference metrics can be initially used. The validation phase has been split into a semantic and a quality validation modules. In the first module, image semantic is taken into account by a human operator to ensure that the image content is coherent with the final task. For example, the image should not be upside down, the subject should be clearly visible and not occluded or that some colors should be in agreement with ideal color classes (like skin, vegetation, sky, etc). As another example, if the image is a photo identification to be printed and included in a passport, the image should satisfy several legal constraints such as: front shot, eyes open, no shadows on the face, etc. These constraints can be ensured by a human analysis or, in very specific cases, by computational algorithms. The second validation module refers to the perceptual quality of the image. In this case NR metrics have to be applied (e.g. bluriness, noise, colorfulness, etc). After the image pass the validation steps, it may undergo a processing phase to make it more suitable for specific task such as enhancement, scaling, compression (e.g. the image must be embedded into a PDF document), etc. Since in these cases we have a source and a processed

image, FR and RR metrics can be used to evaluate the image quality. NR metrics or subjective judgment (either on the processed image only or by comparing the processed and pre-processed images) can be also used. To ensure that the task constraints still hold after the processing, the image should undergo another validation phase. This IQ analysis can be used to give proper feedback in order to improve the IQ of the input before sending it to the printer.

Once the processed image is obtained, it can be sent directly to the printer or, if it is available, to a printer emulator software that taking into account all the characteristics of the HW/SW of the real printer, inks and paper, is able to generate an image of what the print will look like (soft proofing). This soft printed image can be used to estimate the quality of the final printed document using FR, RR, NR metrics or subjective judgements. Finally, the quality of the actual printed image can be assessed according to the specific task. In this case, the evaluation is mainly subjective since it must take into account the print usage (fliers, brochures, art catalogue, high fidelity reproduction, etc.) and possibly the creator intents and preferences. To assess the quality, care should be taken to set up properly viewing condition (light, background, etc).

A similar approach can be used when printing composite documents with several images on a single page. In this scenario, quality can be independently assessed on each image using the above workflow, then a "coherence analysis" could be performed to ensure that, for example, the color features of similar images are in agreement among each other or that all the images belong to a similar semantic class (e.g. indoor, outdoor, landscape, etc).

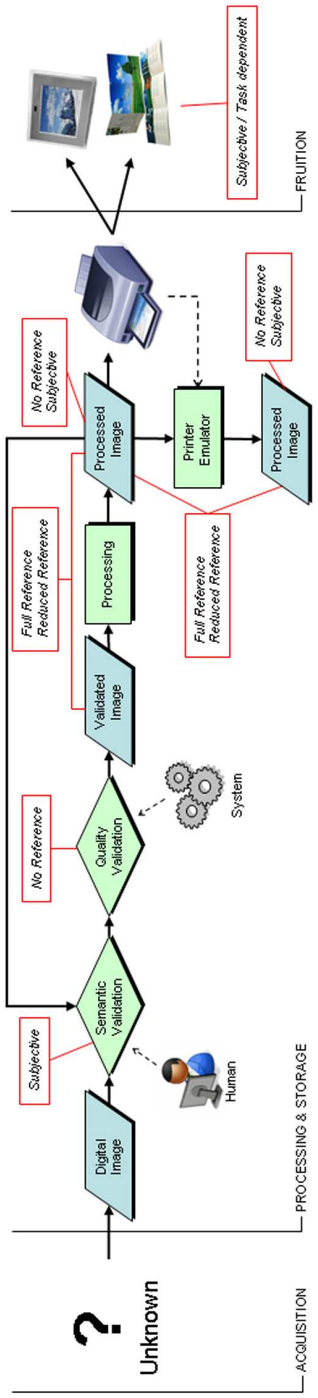


Figure 1.10: A generic image printing workflow chain.

Chapter 2

No-Reference Zipper Metric

The present work concerns the analysis of how demosaicing artifacts affect image quality and proposes a novel no reference metric for their quantification [88]. This metric that fits the psycho-visual data obtained by an experiment analyzes the perceived distortions produced by demosaicing algorithms. The demosaicing operation consists of a combination of color interpolation and anti-aliasing algorithms and converts a raw image acquired with a single sensor array, overlaid with a color filter array, into a full-color image. The most prominent artifact generated by demosaicing algorithms is called zipper. The zipper artifact is characterized by segments (zips) with an On-Off pattern. We perform psycho-visual experiments on a dataset of images that covers nine different degrees of distortions, obtained using three color interpolation algorithms combined with two anti-aliasing algorithms. We then propose our no-reference metric based on measures of blurriness, chromatic and achromatic distortions to fit the psycho-visual data. With this metric demosaicing algorithms could be evaluated and compared.

This chapter is organized as follows. In section 2.1 we briefly describe the demosaicing process, while in section 2.2 we describe how we have generated the dataset utilized during our tests and the psycho-visual experiments that we have conducted to rank the chosen algorithms. From the analysis of the experimental data detailed in section 2.3, we propose our novel no-reference metric described in section 2.4, based on measures of blurriness, chromatic and achromatic distortions. Finally, we report details of the regression we have proposed to fit the subjective data.

2.1 Demosaicing

To produce a color image there should be at least three color samples at each pixel location. The more expensive solution consists in using a color filter in front of each sensor, generating three full-channel color images. Thus, many modern cameras use a color filter array (CFA) in front of the sensor so that only one color is measured at each pixel. This means that to reconstruct the full-resolution image, the missing two color values at each pixel should be estimated. This process, known as demosaicing [55] is generally composed of a color interpolation algorithm followed by an anti-aliasing algorithm to reduce possible artifacts. Among various CFA patterns, the Bayer pattern was the most popular choice [7]. The Bayer array measures the green image on a quincunx grid and the red and blue images on rectangular grids, obtaining 1/2 of the pixels for the green channel, and 1/4 for both the blue and the red channels, as depicted in Figure 2.1.

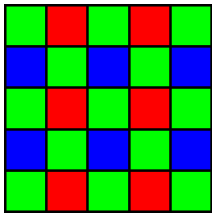


Figure 2.1: The array of filters of the Bayer Pattern.

The most prominent artifact generated by demosaicing algorithms is called zipper. The zipper effect refers to abrupt or unnatural changes of color differences between neighboring pixels, manifested as an "On-Off" pattern [67]. In Figure 2.2 an example of an original image and two different demosaiced versions is reported. As can be seen from Figure 2.2b, the zipper artifacts produced by most of the algorithms are both chromatic and achromatic. On the other hand, demosaicing algorithms that try to mitigate this On-Off pattern, significantly blur the image (Figure 2.2c).

Several algorithms for demosaicing were developed in the literature ([23],[57], [36], [10]), and some of them are proprietary. A survey of these methods was presented by Li et al. [65]. We have here considered nine different demosaicing algorithms obtained combining three color interpolation (CI) algorithms with two anti aliasing (AA) algorithms.

Figure 2.2: a. Original image before demosaicing. b. Demosaiced image (the algorithm adopted here is a combination of the bilinear interpolation and the anti aliasing algorithm porposed by Freeman [36]). The artifacts introduced can be distinguished into achromatic and chromatic zipper. c. Demosaiced image, visibly blurred, obtained by applying an algorithm that tries to mitigate the On-Off pattern (combination of bilinear interpolation and the anti aliasing algorithm proposed by Lu [67]).

The three CI algorithms adopted here are:

- Bilinear interpolation [65]: it is the simplest demosaicing algorithm and acts as a benchmark; the missing values on the three channels are computed by linear interpolation independently.
- ST1: proposed by Smith [106], it performs an isotropic interpolation that includes a non-linear step that minimizes the energy of aliasing artifacts.
- ST2: proposed by Guarnera et al. [41], it uses an elliptic shaped Gaussian kernel to interpolate data, according to the gradient information to better exploit spatial correlation. The authors also included an enhancement step to restore the lost high frequencies.

For what concerns the AA algorithms, we have here considered:

- an algorithm authored by Freeman [36] that suppresses demosaicing artifacts by applying a median filtering to the chrominance channels (R-G) and (B-G) to support the reconstruction of the R and B channels. The red and blue values estimated from the median filtered are used only at pixels where there is no R or B sensor value directly available.

- an algorithm authored by Lu [67] that proposes an anti-aliasing step to extend Freeman's median filtering method by lifting the constraint of keeping the original CFA-sampled values intact.

The nine combinations of these algorithms (summarized in Table 2.1) produce different levels of the typical demosaicing distortions. The choice of these algorithms does not affect the effectiveness of the proposed methodology.

Table 2.1: The nine demosaicing algorithms adopted to obtain the dataset of distorted images.

	Algorithm	Color Interpolation (CI)	Anti-Aliasing (AA)
1	bi	Bilinear	none
2	bifree	Bilinear	Freeman
3	bilu	Bilinear	Lu
4	ST1	ST1	none
5	ST1free	ST1	Freeman
6	ST1lu	ST1	Lu
7	ST2	ST2	none
8	ST2free	ST2	Freeman
9	ST2Lu	ST2	Lu

2.2 Psycho-Visual Setup

2.2.1 Testing Dataset

To perform the subjective data analysis described in this work we have generated a data set of distorted images (which we have called Zipper database) starting from the 24 images of the Kodak photoCD pcd0992 database available at `http://r0k.us/graphics/kodak/`. We have created the mosaiced images by deleting two of the three RGB values at each pixel of the full-color images, and then we have demosaiced them with the nine algorithms of Table 2.1. The database is therefore formed by a total of (24 images x 9 demosaicing methods =) 231 images. The image testing database has been

created to satisfy a good compromise between the number of distortions and the number of different visual contents, keeping in mind that psycho-visual sessions should be limited in time to be reliable. In our work we evaluate the visual impact of the artifacts generated by demosaicing methods, and do not perform a quality evaluation of the algorithms themselves.

2.2.2 Testing Methodologies

For the quality analysis of the images we adopted two different test methods: Single Stimulus method (1S), and Double Stimulus method (2S) [1]. Our goal was to evaluate the perceived quality of the rendered images; for this reason we have chosen to set up a single-stimulus test as our primary source of psycho-visual data, but we were also interested in gathering as much data as possible from the viewers, so we have also conducted a double-stimulus test. We followed Sheikh et al. [103] in setting up our tests by including the original images in both tests and calculating the Difference Score (DS) as the difference between the scores of the original and the distorted image. This way we have obtained different data from different setups with the same unit of measure. In the case of the Single Stimulus method, all the images (rendered images and the original one) are individually shown. While in the Double Stimulus method, the reference image (original image) is shown together with each of its rendered versions. The 1S method can thus be considered as an approximation of the 2S one, as the original image is evaluated only once. The fundamental difference between these two methods is that the Double Stimulus one uses an explicit reference, while the Single Stimulus one does not use any explicit reference.

To perform the psycho-visual tests, the images that have to be judged to obtain a quality rank were shown on a web-based interface (Figure 2.3). A Javascript slider assigning a quality score was used. The workstations adopted were placed in an office environment with normal indoor illumination levels ([103] [5]).

We used five 19-inches CRT COMPAQ S9500 display monitors:

- All the monitors were calibrated with a colorimeter (D65, gamma 2.2).
- Their resolution is 1600X1200 pixels, which corresponds to 110 dpi (using 18 inches as the physical diagonal of the screen as indicated by the manufacturer of the monitors)

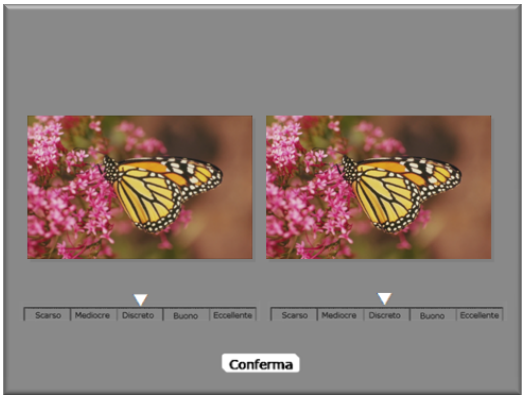


Figure 2.3: The web interface used during the Double Stimuls tests.

- The ambient light levels (a typical office illumination) were maintained constant between the different sessions. There were no reflections on the screens.
- The distance between the observer and the monitors was about 60 cm (corresponding to about 46 pixels per degree of visual angle).
- The refresh rate of the monitors was 75Hz.

In all our experiments distorted images are shown in random order, different for each subject. In the case of the Double Stimulus method the relative position of the original with respect to its distorted version is random in the pair shown.

The panel of subjects involved in this study was recruited from the Psychology Department. The subject pool consisted of students inexperienced with image quality assessment and image impairments. The total number of subjects involved in our experiments is 39, divided into 3 groups as follows: 9 subjects involved in tuning experiments, and 30 subjects involved in 1S and 2S experiments, 15 for each test group.

2.2.3 Psycho-Visual Experiments

In our experiments for the collection of subjective data, we performed three main sessions: a tuning session (where we verified the test efficacy and the

best way to perform the experiment), a preliminary session (where we trained the observers about the nature and the range of the distortion) and a final test session.

The total number of subjects involved in our experiments is 39, divided into 3 groups:

- 9 subjects involved in tuning experiments;
- 15 subjects involved in the single stimulus experiments (both preliminary and test sessions);
- 15 subjects involved in the double stimulus experiments (both preliminary and test sessions).

Note that each subject only belongs to one group. Each subject has been individually briefed about the modality of the experiment in which he has been involved.

All the images utilized for the psycho-visual tests were cropped to fit the dimension of the screen. In particular, to avoid the undersampling of the images used in the Double Stimulus tests, we have cropped all the images to fit a 600 x 600 box, producing respectively images of 600x512 or 512x600. The remaining part of the box has the same color of the background (Figure 3b and c and Figure 4b). Each image has been cropped manually to keep the relevant part of the scene centered, to avoid interferences in the user's judgment, due to a non significant cropping.

Tuning Session

Before starting the preliminary and test sessions, an initial analysis of the test structure and organization was performed to better tune the successive experiments. The 9 subjects participating in this session were not involved in other experiments. During this tuning session we verified the test efficacy and the best way to perform the experiments. In particular, we defined the best visualization time for each image or pair of images on the screen, and the maximum duration of the whole experiment for each participant. We have also collected the following considerations:

- The subjects assume and maintain the correct position and distance from the monitor for the duration of the experiment.

- 30 minutes is the maximum duration of the test for each subject. For longer periods attention decreases and subjects tend to get tired.
- In the case of Double Stimulus test, where the two images are compared, the sliders and the quality scales must appear contemporarily on the screen.

Regarding comments and considerations of the subjects involved in this tuning session, we have determined the minimum time of image visualization that permits an appropriate quality evaluation.

Preliminary Session

During a preliminary test, each subject was implicitly trained about the nature of the distortion he was going to evaluate. In particular, he was trained about the range of the distortion intensity. These preliminary sessions were necessary to avoid this training phase during the effective test, thus conditioning the experimental results. We had preliminary sessions for all the subjects involved (except for 9 subjects involved in the tuning phase) and for each of the experiments (1S and 2S). Thus we had preliminary sessions for all the subjects involved and for each of the experiments: Double Stimulus and Single Stimulus. Four images were chosen from the entire database. The demosaicing algorithms applied to these images where the Bilinear and the ST proprietary. We have decided to apply these two algorithms because they were supposed to be the worst and the best ones. In this way the subjects experience the entire distortion range before starting the effective test.

For the test session we used 10 images from the 24 of the original database, together with their corresponding 9 distorted versions, (for a total of 100 images). The 10 images chosen for this session are shown in Figure 2.4. Note that we had to keep the number of analyzed images limited to 100, since subjects can pay attention only for up to 30 minutes. After this time their judgments are no longer reliable [1]). The number of test images is however aligned with what is done in the literature. In the work of Nyman et al. [77] for example, 9 image processing pipes applied to 8 image contents (for a total of 9x8=72 test images) were evaluated with a psycho-visual experiment involving 14 test subjects. In other works that involve psycho-visual experiments, the number of original images considered is even lower, four images each printed on 15 different papers [62], or five images each with 15 different levels of sharpness [85]. The greater the number of algorithms/processing to

be evaluated, the lower the number of original images that can be considered to keep the time of the experiment reasonable.

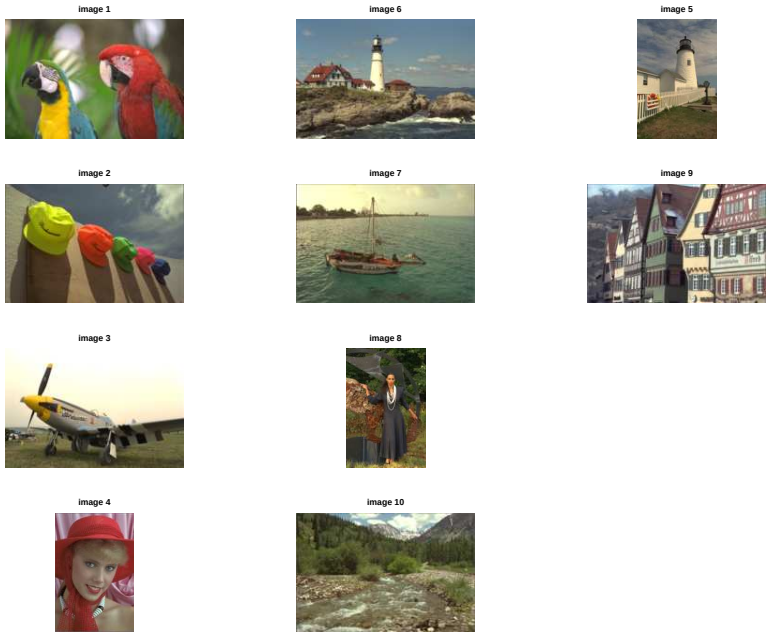


Figure 2.4: The 10 original images utilized during the test session.

2.2.4 Data Processing

As the different algorithms considered produce different levels of the typical demosaicing defects (chromatic and achromatic zipper, blur) we analyzed the subjective evaluation of these defects through the subjective rank of the algorithms. The data processing here described is applied for both the test methods (1S, 2S) adopted for collecting the experimental data. For each subject j-th and distorted image i-th we evaluated the perceptual distance between original and distorted images in terms of difference of assigned scores (Difference Score, DS):

$$DS_{ij} = So_{ij} - Sd_{ij} \tag{2.1}$$

where Sd_{ij} represents the score assigned by the j-th subject to the i-th distorted image, while So_{ij} the score of the reference image corresponding to the i-th distorted image; $j = 1, \ldots, J$, denotes subjects belonging to the group of J individuals, and $i = 1, \ldots, S \times T$ denotes the distorted image, with S number of reference images, and T number of algorithms to be evaluated. For each subject we evaluated the standard-DS_{ij} (zDS_{ij}), a DS distribution normalized with respect to the subject [103], as:

$$zDS_{ij} = \frac{(DS_{ij} - M_j)}{\sqrt{V_j}} \tag{2.2}$$

where $M_j = \frac{1}{S \times T} \sum_i^{S \times T} DS_{ij}$, $V_j = \frac{1}{S \times T} \sum_i^{S \times T} (DS_{ij} - M_j)^2$ are respectively the mean and variance of DS_{ij} with respect to the j-th subject. For each algorithm $t \in T$, we evaluated the final score R_t by summing zDS_{ij} of equation 2.2 over the subjects $j \in J$, and on the reference images $i \in S$.

$$R_t = \frac{1}{J \times S} \sum_{j \in J} \sum_{i \in S} zDS_{ij} \tag{2.3}$$

The rank of the algorithms is then obtained sorting these final scores. We also calculated the rank of the algorithms starting from the median with respect to subject $j \in J$ and reference images $i \in S$.

$$MR_t = median(zDS_{ij}) \tag{2.4}$$

2.3 Data Analysis

In analyzing distorted images supposed to be worse than the original, we expect all the DS values (distance between the scores of the original image and the rendered image) to be positive. It happened sometimes in our experiments that distorted images were judged better than the corresponding original. This phenomenon is called inversion. We have decided to maintain all the inversions.

2.3.1 Inversions

We define Just Noticeable Difference Threshold (JND) the threshold under which differences between distorted images and their original are not noticeable. Assuming that this threshold exists, the inversions can be classified into three categories:

JND Inversions The subject is not able to distinguish between the original and the distorted image. The inversion is unintentional.

Preferential Inversions The subject prefers the elaborated image.

Error Inversions The subject does not properly use the interface and in particular, assigns a wrong value in the quality scale.

As reported in [1], the inversions are usually handled, following a standard procedure:

1. The JND threshold is estimated with a Pairwise Comparison (PC) test [111];
2. Inversions that produce values under the JND threshold (JND Inversions) are taken into account in the final analysis;
3. Inversions that produce values over the JND threshold are considered as Error Inversions. Their absolute values are taken into account in the final analysis.

Preferential Inversions

In [66], the authors report interesting considerations about Preferential Inversions in case the of images processed by demosaicing algorithms. They analyze the results of a Pairwise Comparison test of images processed by different demosaicing algorithms. This psycho-visual experiment demonstrates that certain algorithms produce distorted images judged better that the original. This preference is due to the apparent sharpness introduced by these algorithms. It is well known that sharpness plays an important role in the evaluation of apparent quality of digital images, [52] [56]. Using a Double Stimulus method as the PC test, the original image appears blurred in comparison with the elaborated one. Not all the demosaicing algorithms analyzed

in our experiment show the same sharpening behavior. As a consequence, the collected data are non-homogeneous with respect to algorithms that present different levels of Preferential Inversions. Applying the standard procedure, Preferential Inversions are not explicitly considered. These Inversions fall both in the Error inversions and in the JND Inversions. For this reason we have decided to maintain all the inversions. This decision requires the solution to two different problems:

- How to treat the Error Inversions?

  The Error inversions cannot be common to different subjects. They are anomalous values with respect to the score distribution of each algorithm. We are not interested in finding the Error Inversions; we just would like to verify that they do not alter the data analysis. To this end we have validated the final rank of the algorithms with the analysis of the median of the Difference Score, which is a more robust measure with respect to noise.

- How to treat the Preferential Inversions?

  Maintaining the Preferential Inversions, the DS measure cannot be further considered as a distance between the reference image and the distorted one with respect to the analyzed artifact (zipper artifact), as we have previously discussed. The influence of these inversions appears to be different in the case of Single Stimulus and Double Stimulus tests. In fact, the effect of the introduced sharpness is lower in the case of the 1S test because there is not a simultaneous comparison with the original image. Thus, the analysis of the 1S test results with respect to the 2S ones can be useful for evaluating this phenomenon.

2.3.2 Features Identification

We want to emphasize that with this data analysis we are not evaluating the performance of the algorithms, but instead we are interested in highlighting the major effects that influence the subjective evaluations of the perceived quality of demosaiced images. The final goal is to identify the significant features to be used in a proper metric so that it can be able to reproduce the experimental data. In Figure 2.5, the rank of the nine demosaicing algorithms obtained combining the three color interpolation (CI) algorithms with the two anti-aliasing (AA) algorithms listed in Table 2.1 are reported

for both the 1S and the 2S experiments. Figure 2.5(a) and Figure 2.5(b) show the ranks of the 2S experiment using respectively the mean R_t and the median MR_t as a central tendency indicator. The coherence between these two ranks confirms the stability of the results. In Figure 2.5(c) and Figure 2.5(d), the same data are reported for the 1S experiment. In Figure 2.6 a comparison of the two experiments is reported. The solid line refers to the Single Stimulus (1S) experiment, while the dotted line refers to the Double Stimulus (2S) experiment. As a preliminary step, we have grouped the 9 demosaicing methods into triplets, with respect to the CI algorithm applied. As a general consideration, CI algorithms alone (i.e. bilinear, ST1 and ST2) were judged worse than their corresponding versions coupled with any of the AA algorithms considered. With respect to the CI approach, the ST2 method (coupled with any AA algorithm) is always preferred as it produces sharper images. This behavior is due to the explicit boosting introduced by the authors to restore the lost high frequencies. These results confirm that sharpness plays an important role in influencing image quality judgments, [56] and [52]. 1S tests are less precise than 2S tests because the reference image is shown only once, and the comparison between distorted images and reference ones is more difficult. Were this the only difference between the two tests, we would not expect significant changes in the algorithm ranks. This assumption was not fully verified in our experiments. This discrepancy is also due to the effect of the perceived sharpness on image quality, which is more evident in 2S tests due to the direct comparison with the reference images. The AA algorithms considered have influenced the image sharpness at different degrees. In particular the Freeman algorithm produces a sharper image, while the Lu algorithm makes the images more blurred. This phenomenon is more evident when these anti-aliasing algorithms are coupled with the basic color interpolation method (bilinear interpolation) as shown in Figure 2.7. As a consequence, the rank positions of the algorithms labeled as bifree (algorithm 2) and st2free (algorithm 8) are swapped from the 2S to the 1S experiment with respect to the corresponding bilu (algorithm 3) and st2lu (algorithm 9) as shown in Figure 2.7.

2.3.3 Image Frequency Content

We have analyzed the experimental data with respect to the image frequency content to investigate the cross-talks between the zipper artifacts introduced by the color interpolation process and the image frequencies.

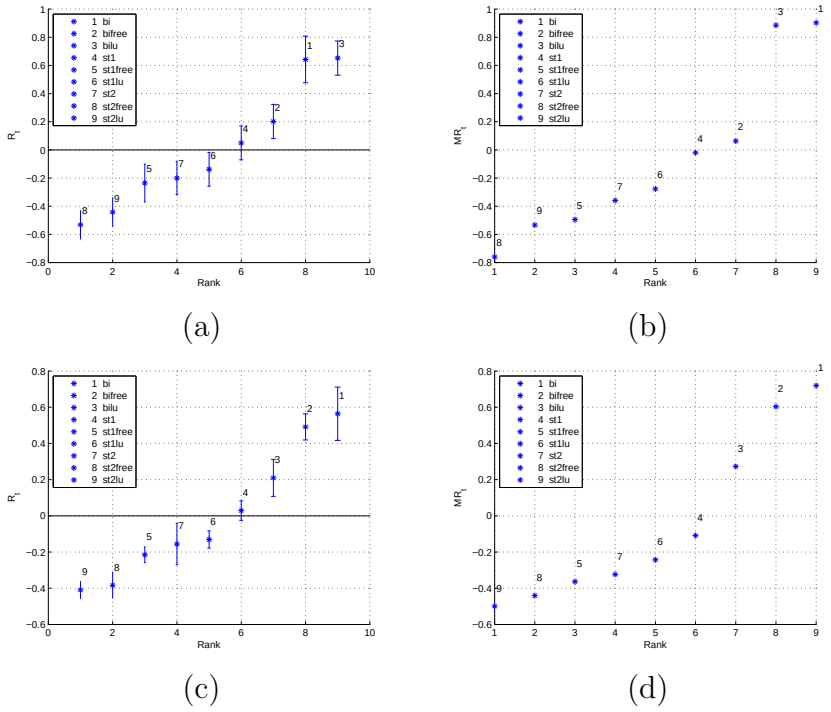


Figure 2.5: (a) Algorithm ranks in terms of final score R_t in the 2S experiment. (b) The 2S rank resulting from using the median MR_t as a central tendency indicator [103]. (c) Algorithm ranks in terms of R_t in the 1S experiment. (d) The 1S rank resulting using the median MR_t as a central tendency indicator

In Figure 2.4, the 10 images used in our tests are roughly separated so that: the first column reports images with few details, Low-Frequency (LF) set; the second column shows images with fine details, Middle-Frequency (MF) set; while in the third column two images of High Frequency (HF) are depicted.

To better understand how the frequency content influences the psychovisual data, we have collected the subjective score ($Score_i$) for each of the ($S = 10$) test images and for each of the ($T = 9$) demosaicing algorithms. Summing the zDS_{ij} of equation 2.3 over the subjects $j \in J$ we obtain:

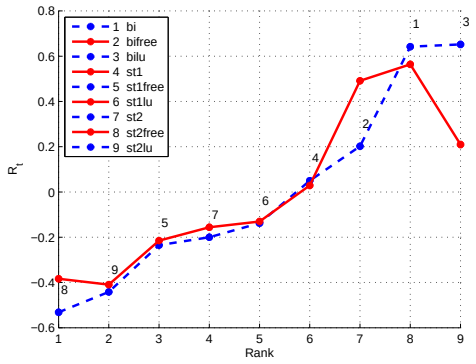


Figure 2.6: Comparison between R_t values of the 1S experiment (red solid line) and the R_t values of the 2S experiment (blue dotted line).

$$Score_i = \frac{1}{J} \sum_{j \in J} zDS_{ij} \tag{2.5}$$

with $i = 1, ...S \times T$. The $Score_i$ are reported for both 2S and 1S experiment in Figure 2.9, where the layout of Figure 2.4 is maintained. In particular, the first column corresponds to the LF set, the second column to the MF set, while the last column corresponds to the HF set. Each subplot reports the experimental $Score_i$ corresponding to the 9 distortions applied to each image. These scores are grouped into triplets with respect to the CI method (bilinear + three AA, ST1 + three AA, and ST2 + three AA).

In the following analysis we have decided to eliminate the two images of the HF set. These images are characterized by a texture with a near-Nyquist frequency as shown in Figure 2.8, where the distortions due to the aliasing are evident. Images belonging to this HF set have suffered from very strong distortion after the color interpolation process and thus their subjective scores could have been influenced by the near-Nyquist artifacts, which are not object of our study.

We can notice that images with a comparable level of details share common patterns in their scores. In particular, when the achromatic zipper (mainly produced by the Freeman AA algorithm) is combined with middle-high frequency content, not only the contrast of the zipper highlights the edges, but also the middle-high frequency content masks the On-Off pattern.

(a) (b) (c)

Figure 2.7: Detail of an image rendered with different algorithms (a) Bilinear (CI) (b) Bilinear (CI) + Freeman (AA) (c) Bilinear (CI) + Lu (AA)

This combined effect results in a sharper appearance of the image; this is more evident when the images are directly compared with the reference, as in the 2S test. This behavior is related to the texture masking effect of the human visual system [80]. From the point of view of the algorithm ranks (Figure 2.9), these considerations are confirmed by the good performance on the MF set obtained using the Freeman AA with respect to each triplets of CI algorithms (algorithms number 2, 5, 8). On the other hand, when the algorithms that produce this achromatic distortion are applied to images of the LF set, the high contrast of the zipper pattern and its On-Off structure remain visible. In fact, the evaluation of the CI algorithms coupled with the Freeman AA in the case of LF set is worse than in the case of the MF set, especially in the case of the 1S experiment where the sharpness is less perceived.

For what concerns the chromatic zipper, the behavior is simpler. This artifact is more visible as the number of edge pixels in the image increases, and it seems to be immune to masking effects. For this reason we chose to discriminate between chromatic and achromatic distortion.

2.4 No-Reference Metric for Demosaicing

The data analysis confirms that the perceptual quality of demosaiced images depends on sharpness, and on chromatic and achromatic zipper. For this reason we have decided to define our no-reference metric considering the

following three aspects separately:

Blur as index of lack of sharpness. The corresponding measure is indicated as B in what follows.

Chromatic zipper distortion (measure indicated as CD)

Achromatic zipper distortion (measure indicated as AcD)

Thus, the demosaicing metric DM that we have developed is composed of three properly scaled terms, corresponding to these three aspects:

$$DM = B + CD + AcD \tag{2.6}$$

We chose a sum expression because when one of these terms is significantly high, the others are less significant. This consideration arises from the experimental evidence of the behavior of different demosaicing algorithms. A strong low pass filtering adopted to reduce the zips produces a blurred image, and thus in this case the blur measure B is dominant with respect to the others. In case of more conservative filtering, the image sharpness is preserved, but the zips still remain as a defect. Different color interpolation algorithms produce zips with different levels of saturation, ranging from achromatic to highly saturated zips.

2.4.1 Blur

The blur in an image is due to the attenuation of the high spatial frequencies. Blur is the typical artifact in out-of-focus shots, but it may also be caused by the relative movement between camera and subject (motion blur), and by the encoder (compression blur). In the context of color interpolation artifacts, blurriness is due to an excessive low pass filtering of the anti-aliasing algorithm. Marziliano et al. [72] present a blind (no-reference) blur metric that is based on measuring the average spread of the vertical edges. They define the edge spread as the distance between the local minima (p_1) and the local maxima (p_2) nearest to the edge along the gradient direction(Figure 2.10). The edge spread was used to predict the quality of jpeg2000 compressed images and has shown to be consistent with the observers' ratings obtained in subjective experiments. We use as blur indicator, the average edge spread of the image **Es**, evaluated as follows:

$$\mathbf{Es} = \frac{1}{NEdge} \sum_{e \in Edge} dist_4(p_1, p_2) \tag{2.7}$$

where $Edge$ is the set of edge pixels of the image and $NEdges$ is the number of these edge pixels. We chose to estimate the edge spread by searching around the edge in four directions (indicated with $dist_4$ in 2.7) horizontal, vertical, +45 and -45 degrees.

2.4.2 Chromatic and Achromatic Zipper

The zipper pattern detection was carried out as follows. On the gray-scale image, we computed the gradient magnitude in both directions with the following convolution kernels:

$$\begin{aligned} V &= \begin{bmatrix} -1 & 1 \end{bmatrix} \\ H &= \begin{bmatrix} -1 & 1 \end{bmatrix}^T \end{aligned} \tag{2.8}$$

The two gradient maps, G_x and G_y (horizontal and vertical), are treated separately to detect zipper segments. Working on the horizontal direction, we first compute the gradient sign map by quantizing the gradient magnitude as follows (the same process is extended to the vertical direction):

$$SignMap_x(x, y) = \begin{cases} 2 & \text{if } G_x(x, y) < 0 \\ 1 & \text{if } G_x(x, y) > 0 \\ 0 & \text{if } G_x(x, y) = 0 \end{cases} \tag{2.9}$$

Thus, a zipper segment (which is an On-Off pattern) is characterized in the sign map by a sequence of alternating 2s and 1s (see Figure 2.11 (b)). The number and the extension of zips is not sufficient to quantify the perceived quality of a color interpolation algorithm. In fact, some zipper pixels are more visible than others (see Figure 2.11 (c)).

To evaluate the visibility of the pixels belonging to the zipper segments, we compute $DL(x, y)$ and $DC(x, y)$ distances between adjacent pixels in zipper segments, starting from the CIE-94 definitions [97]:

$$\begin{aligned} DL(x,y) &= ((\Delta L^*(x,y))^2)^{\frac{1}{2}} \\ DC(x,y) &= ((\Delta C^*(x,y)/S_c)^2 + (\Delta H^*(x,y)/S_H)^2)^{\frac{1}{2}} \end{aligned} \tag{2.10}$$

where [1]

$$\begin{aligned} \Delta L^*(x,y) &= L^*(x,y) - L^*(x,y-1) \\ \Delta C^*(x,y) &= ((a^*(x,y))^2 + (b^*(x,y))^2)^{\frac{1}{2}} - ((a^*(x,y-1))^2 + (b^*(x,y-1))^2)^{\frac{1}{2}} \\ \Delta H^*(x,y) &= ((\Delta E_{76}(x,y))^2 - (\Delta L^*(x,y))^2 - (\Delta C^*(x,y))^2). \end{aligned} \tag{2.11}$$

We calculated the median of $DL(x,y)$ with respect to the whole set of zipper segments in both directions and averaged them. We performed the same calculations for $DC(x,y)$, obtaining two indicators labeled as **DL** and **DC** in what follows. These two indicators, together with the average edge spread (**Es**) and the percentage of zipper pixel in the image (ZpA), were used to calculate the overall metric.

2.4.3 Metric Parameter Estimation

Starting from the the blur and zipper pattern analysis described in the two previous subsections, our demosaicing metric (2.6) can be rewritten as:

$$DM = w_B \times \mathbf{Es} \times + w_C \times \mathbf{DC} + w_L \times e^{\mathbf{DL-DC}} \times ZpA \tag{2.12}$$

Algorithms that reduce aliasing tend also to desaturate the zips, increasing the coherence between channels. This effect produces achromatic zips, where **DL** exceeds **DC**. w_B, w_C and w_L are weights to be chosen so that our metric can predict the algorithms' rank produced by the psycho-visual experiments. To this end, we have applied the proposed metric to the images in the Zipper Database, and then we have calculated the average scores of

[1] The equations are reported only for the horizontal case. In the calculation of the differences we excluded the non-zipper pixels. ΔE_{76} is the standard Euclidean distance between the $L^*a^*b^*$ coordinate of the adjacent pixels.

the nine algorithms. We have found the regression functions that permit the best fit between the average values given by our measure and the average subjective responses for both the 2S and 1S data. These fittings are reported in Figure 2.12.

The contribution of each feature adopted in our metric, can be investigated by looking at the different values assumed by the weights w_B, w_C and w_L in Equation 2.12. These values are reported in Table 2.2. The main difference between 2S and 1S is in the contribution of the sharpness. In fact, in the case of 2S when a reference image is shown, the difference in sharpness is more evident, thus the corresponding weight w_B is higher than in the 1S.

Table 2.2: Weights for the 1S and 2S test data.

Weights	2S	1S
w_B	5.0	2.0
w_L	5.0	5.0
w_C	1.5	1.2

Figure 2.8: Details of the images of Figure 2.4 with near-Nyquist frequency content.

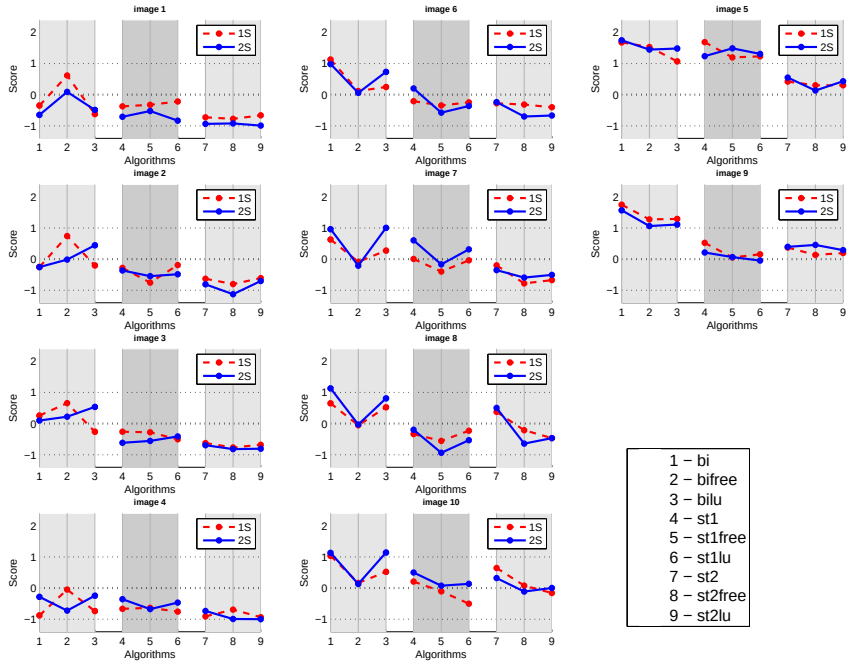


Figure 2.9: Subjective test data. Each subplot refers to the corresponding image of Figure 2.4. The scores are grouped in triplets, corresponding to each of the three color interpolation methods coupled with the three different anti-aliasing strategies. For instance, the first triplet corresponds to algorithms 1,2 and 3, i.e. bilinear interpolation with no anti-aliasing, Freeman anti-aliasing and Lu anti-aliasing respectively.

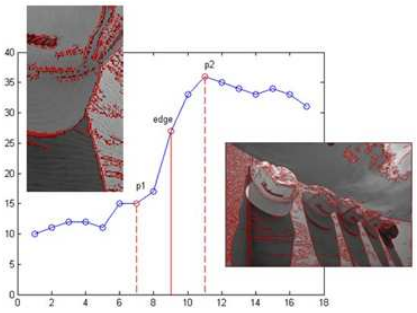
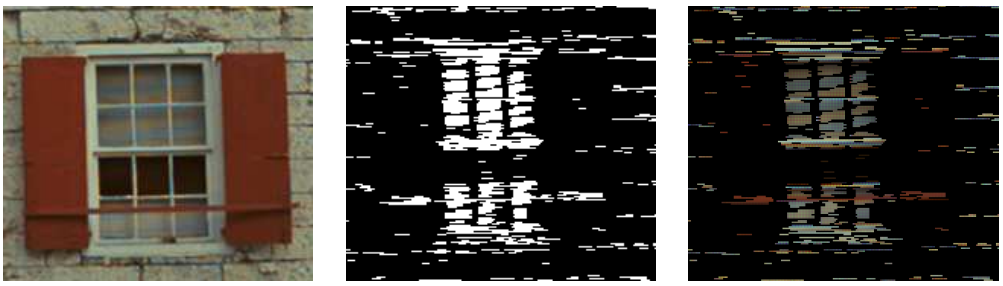
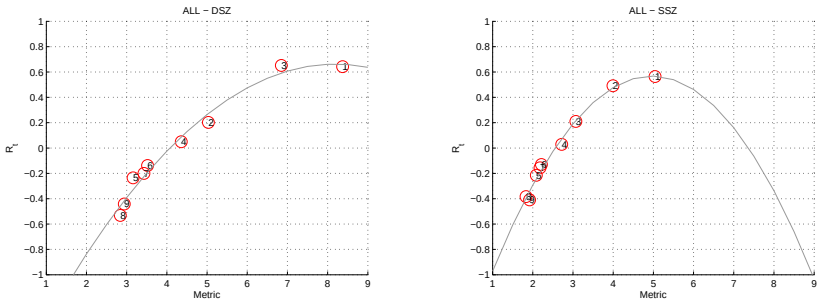


Figure 2.10: Edge spread defined as the distance between the local minima (p_1) and the local maxima (p_2) nearest to the edge.



(a) (b) (c)

Figure 2.11: (a) Detail of an image rendered with bilinear interpolation. (b) Horizontal zipper map. (c) The original image masked with the horizontal zipper map.



(a) (b)

Figure 2.12: Average values of the color interpolation algorithms. Metric versus average subjective algorithm rating. (a) Double Stimulus test data. (b) Single Stimulus test data.

Chapter 3

No-Reference JPEG Metric

No-reference quality metrics estimate the perceived quality exploiting only the image itself. Typically, no-reference metrics are designed to measure specific artifacts using a distortion model. Some psycho-visual experiments have shown that the perception of distortions is influenced by the amount of details in the image's content, suggesting the need for a "content weighting factor." This dependency is coherent with known masking effects of the human visual system. In order to explore this phenomenon, we setup a series of experiments applying regression trees to the problem of no-reference quality assessment [89]. In particular, we have focused on the blocking distortion of JPEG compressed images. Experimental results show that information about the visual content of the image can be exploited to improve the estimation of the quality of JPEG compressed images.

3.1 Overview

Image quality metrics are designed to estimate the perceived quality of images. A perfect metric (e.g. a metric that takes into account the HVS sensibility to the distortion) would be linearly related to the Mean Opinion Score (MOS). However, it is likely that the parameters of the linear model would depend on the content of the images. In fact, some images are more "suitable" for the distortion to be seen: the same amount of distortion will be perceived differently on different categories of images (e.g. clouds in the sky and a shot of a building). Our approach is based on the assumption that the content of the image alters the parameters of the model. A regression tree

is used to partition the images into clusters characterized by similar content. Then, for each cluster a specific model is fitted to map the metric to the MOS.

3.2 Classification Methodology

The tree growing algorithm is inspired by the Classification And Regression Trees [12] (CART) methodology.

These are binary trees produced by recursively partitioning the predictor space, each split being formed by conditions related to the predictor values. Each subset corresponds to a node of the tree: the whole predictor space corresponds to the root node, the subsets of the final partition correspond to the terminal nodes. Once a tree has been constructed, a class is assigned to each of the terminal nodes, and it is this that makes the tree a classifier: when a new case is processed by the tree, the class associated with the terminal node in which the case ends up on the basis of its predictor values is its predicted class.

In problems where it is feasible to assume that the cost of misclassifying a class j case as a class i case is the same for all $i \neq j$, $i, j = 1, \cdots, J$, the class assigned to each terminal node t is the class i for which $p(i|t) = \max_j p(j|t)$, where $p(j|t)$ is the resubstitution estimate of the conditional probability of class j in node t, that is, the probability that a case found in node t is a class j case. With this rule the resubstitution estimate of the accuracy inside the node, given by $p(i|t)$, is maximized or, equivalently, the resubstitution estimate of the misclassification probability inside the node, given by $1 - p(i|t)$, is minimized. If the prior probabilities of the classes are estimated from the data, $p(i|t)$ is simply the proportion of class i cases inside node t and the resubstitution estime of the accuracy inside the node is reduced to the relative proportion of cases in the node that belong to class i.

When it is not realistic to assume equal misclassification costs, the class assigned to each terminal node of the tree is the class for which the estimated misclassification cost inside the node is minimized. In our study we have assumed equal misclassification costs.

The critical problems of the splitting process are essentially two: how to identify candidate splits, and how to define the goodness of the splits. Candidate splits are generated by a set of admissible questions regarding the

values of the predictors, which differ according to the nature of the predictors themselves. At each step of the process, all the predictors are searched one by one, and the best split, in the sense defined below, is found for each predictor. The best splits are then compared, and the best of these selected.

The idea central to the goodness of splits is that of selecting the splits so that the data in the descendant nodes are purer than the data in the original ones. To do so, a function of impurity of the nodes, $i(t)$, is introduced, and the decrease in its value produced by a split is taken as a measure of the goodness of the split itself.

The function of node impurity we have used is the Gini diversity index

$$i(t) = \sum_{i \neq j} p(i|t)p(j|t) = 1 - \sum_{j} p^2(j|t), \tag{3.1}$$

which has a clear interpretation in terms of variances of Bernoulli variables. If, for each class j, we consider the random variable Y_j , which is 1 (success) if a case of t belongs in class j and 0 (failure) otherwise, it can be modeled as a Bernoulli variable whose probability of success is estimated by $p(j|t)$, and the quantity

$$1 - \sum_{j} p^2(j|t) \tag{3.2}$$

is the sum of the estimated variances of such variables.

In CART methodology the size of a tree is treated as a tuning parameter, and the optimal size is adaptively chosen from the data. A very large tree is grown and then pruned, using a cost-complexity criterion which governs the tradeoff between size and accuracy, or cost. This eliminates both the risk of large trees which overfit the training data, as well as that of small trees that do not capture important information. The pruning process generates a sequence $\{T_l\}_{l \in \{1,\cdots,L\}}$ of subtrees decreasing in size; these are evaluated in terms of their accuracy, or misclassification cost, and the best subtree is then selected. When large sets of data are available, as is the case here, the accuracy, or misclassification cost, of the subtrees are usually estimated on the basis of a test set. Otherwise, cross-validation must be applied.

*A*lthough the pruning process prevents the danger of trees too tailored to the training data, there is still overfitting due to instability, a phenomenon inherent in the hierarchical nature of the construction process of trees. Even a small change in data may result in a very different series of splits, and this clearly affects both the structure of the trees, and the consequent classification results.

3.3 Content Descriptors

Information about the textures and structures within the image can be obtained using a wavelet decomposition. This technique is often used in content-based retrieval for similarity retrieval,target search, compression, texture analysis,biometrics, etc. [44, 96]. In multi-resolution wavelet analysis, at each level of resolution (i.e.at each application of the wavelet decomposition) we have four bands containing different information obtained by applying a combination of a low-pass filter(L) and a high pass filter(H). Specifically, the information corresponds to a low-pass filtered version of the processed image (LLband), and three bands of details that roughly correspond to the horizontal edges (LHband) of the original images, the vertical edges (HLband) and the diagonal edges (HHband). Each band is a matrix of values, one fourth the size of the original image. Wavelet decomposition is applied recursively to the LL band. The resultant decomposition will contain information, i.e.details, at the lower resolution. The process can be repeated until the LL sub-band cannot be further processed or until a given number of wavelet decomposition applications is reached. Different filters can be used to produce the bands of the wavelet analysis [73] e.g. Harr, Daubechies, Symlet, Biort, etc. For our purposes the wavelet statistics features are extracted from the luminance image using a three-iteration Daubechies wavelet decomposition, producing a total of 10 bands. The mean and variance of the absolute values in each band are then computed as band statistics. These feature values represent the energy i.e. the amount of information within each band and provide a concise description of the image's content. This feature thus composed of 20(two energy values for each of the 10 bands) components.

3.4 Proposed approach

The tree is produced by recursively partitioning the set of images, represented by the feature vectors $T = \{\mathbf{f}_1, \ldots, \mathbf{f}_N\}$, $\mathbf{f}_i \in \mathbb{R}^D$ labeled with the corresponding MOS $\mu_1, \ldots, \mu_N$, $\mu_i \in \mathbb{R}$, and the values of the quality metric considered $\{y_1, \ldots, y_N\}$, $y_j \in \mathbb{R}$. The partitioning is driven by an impurity function which measures how well, given a set of images, the relationship between the metric and MOS can be described by a function chosen from a given parametric model M_θ. The impurity $I(S)$ of a non-empty set of images

$S \subseteq \{1, \ldots, N\}$ is defined as:

$$I(S) = \frac{1}{|S|} \sum_{i \in S} (M_{\hat{\theta}}(y_i) - \mu_i)^2, \tag{3.3}$$

where $M_{\hat{\theta}}$ is the function (defined by the parameters $\hat{\theta}$) obtained by fitting the metric to the MOS by a least squares regression.

The recursive tree growing procedure starts to consider the whole set of images. To partition the set P (the parent node, in tree terminology) into two subsets L and R (children nodes) the algorithms consider all the possible splits defined by thresholding the values of the components of the feature vectors. Among all the components and the threshold values, the pair (j^*, τ^*) which maximizes the decrease in impurity is selected:

$$\Delta I(j, \tau) = I(P) - \frac{|L|}{|P|} I(L) - \frac{|R|}{|P|} I(R), \tag{3.4}$$

where L and R are defined according to the split (j, τ): $L = \{i \in P : f_{ij} \leq \tau\}$, $R = P \setminus L$. The optimal pair (j^*, τ^*) is found by an exhaustive search on all possible values of j and τ. To avoid inaccurate estimation of the parameters of the model, the growing process is not applied to small nodes (less than five images in the current setup). Finally, each terminal node is labeled with the parameters θ computed by the least squares regression.

Given a new image the tree determines in which terminal node it falls in on the basis of the values of the feature vector. Then, the corresponding parameters θ are used together with the value of the metric y to estimate the MOS as $M_\theta(y)$.

Each tree has been pruned using the Minimal Cost Complexity Pruning Algorithm [12]. We adopted a k-fold strategy to build the test sets and the training sets. In each of the 29 training sets used, all the versions of one of the 29 original images have been excluded to avoid data snooping.

As content descriptor we chose the mean and the variance of a 3 level wavelet transform, for a total of 20 features (2 indices $\times$ 10 bands) [20]. These features are quite stable with respect to the introduction of JPEG compression. In order to verify this stability, we have tagged each of the original images with a class label. Then we trained a classification tree (with the CART algorithm) using the wavelet features to predict the class of each of the 175 image used. The error (in the non-pruned tree) is zero, which means that the wavelet features are able to discriminate between the

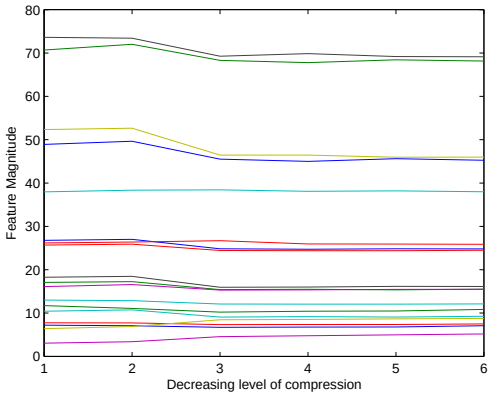


Figure 3.1: Compression invariance of the wavelet features. For the 6 versions of the "womanhat" image we plotted the magnitude of the 20 components of the features (one for each line).

different contents independently of their level of compression. As shown in Figure 3.1 for the "womanhat" image, the wavelet features are robust with respect to different levels of compression.

For MOS estimation, we empirically chose a logarithmic model. This model has the advantage of being monotone and is able to catch the log-like relation between metrics and MOS that we observed in different metrics. This behavior can be explained by assuming that the higher levels of distortion proposed to the subjects are beyond their level of "saturation".

$$\hat{\mu} = \theta_1 \log(y + \theta_2^2) + \theta_3, \tag{3.5}$$

where $\hat{\mu}$ is the estimated MOS and y is the value of the metric. To simplify the non-linear regression in each node we normalized the metric values as follows:

$$y' = \frac{y - (y_{min} - \epsilon)}{y_{max} - y_{min}}, \tag{3.6}$$

where y_{min} and y_{max} are respectively the minimum and the maximum of the metric on the whole database (also using the undistorted images), and $\epsilon = e^{-36}$ is a constant that prevents the application of the logarithm to non-positive arguments.

The proposed method is summarized in Figure 3.2.

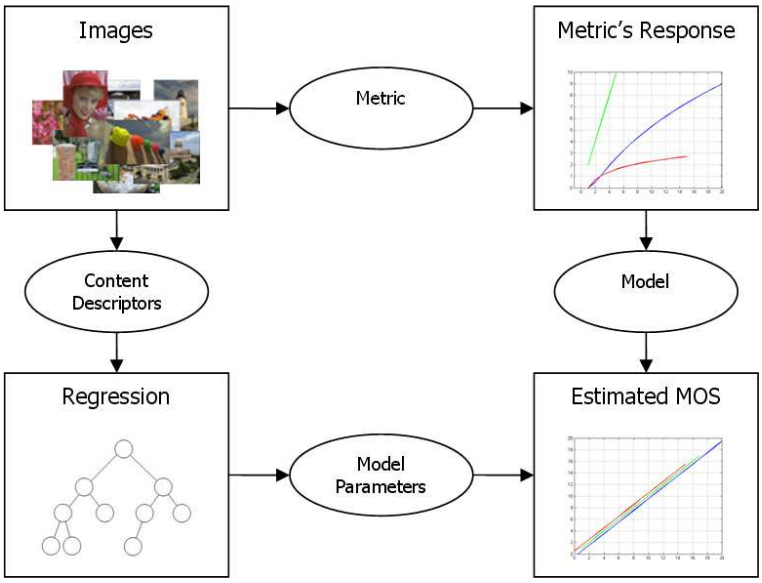


Figure 3.2: The proposed method to embed content information in image quality metrics.

3.5 Experimental Results

For the experimentation we used the JPEG subset of the LIVE database[99]. The database is derived from a set of 29 different color images which have been distorted by JPEG compression. The level of the distortion has been modulated to produce images at a broad range of quality, from imperceptible levels to high levels of impairment. The database contains a total of 175 images with bit-rates ranging from 0.15 bpp to 3.34 bpp. Each image has been evaluated by a mean of 22 human subjects, utilizing a single-stimulus test methodology [1] in which the original images were included; this way it was possible to derive a quality difference score for each image. The authors have made the Difference Mean Opinion Score (MOS) available for each image included in the database. For details, see [101].

We applied our method to four different blocking metrics (Table 3.1).

For each metric the method has been evaluated by comparing the estimated MOS with the correct one. The average mean square error (MSE) has been computed as a measure of error.

For the sake of comparison, for each metric we also computed a global

Table 3.1: The metrics analyzed during our tests with their main characteristics. For further details, see the references.

Name	Application	Method	Other aspects considered
WBE [122]	JPEG	Magnitude of the blocking signal in the frequency domain	
WSB [124]	JPEG	Magnitude of the blocking signal in the spatial domain	Signal activity correction
PAN [79]	JPEG	Magnitude of the blocking signal in the spatial domain	Flat area correction (for a very low Q-factor)
VLA [118]	MPEG	Ratio between intra-block similarity and inter-block similarity	

Table 3.2: Mean Square Error (MSE) obtained by a global regression (using the log model) and the proposed method.The number of leaves refers to the pruned trees trained for each metrics.

	MSE		
Metric	Global regression	Proposed method	Number of leaves
PAN	89.74	26.90	8
VLA	90.86	50.40	6
WSB	60.48	23.71	7
WBE	47.00	18.41	7

regression on the whole dataset using the same logarithmic model. The results obtained are reported in Table 3.2. Figure 3.3 reports the results obtained for the PAN metric. The first plot is the global regression result; the second one is the result with the proposed method. The third plot shows the different instances of the model that were used. The two bottom maps show the distribution of the images in the different leaves of the pruned tree (the entire distribution on the left map and the mode on the right). In the same way Figures 3.4, 3.5, and 3.6 report the results obtained with the VLA, WBE and WSB metrics.

The best results have been obtained with the WBE metrics. This metric shows some regularities that make it easy to identify a good regression model.

It is also one of the simplest metrics: it is designed to measure the blocking effect (discontinuities at block boundaries) ignoring other aspects (e.g. flat blocks at a high level of compression, or the activity of the signal). Good predictions were also obtained with the WSB metric.

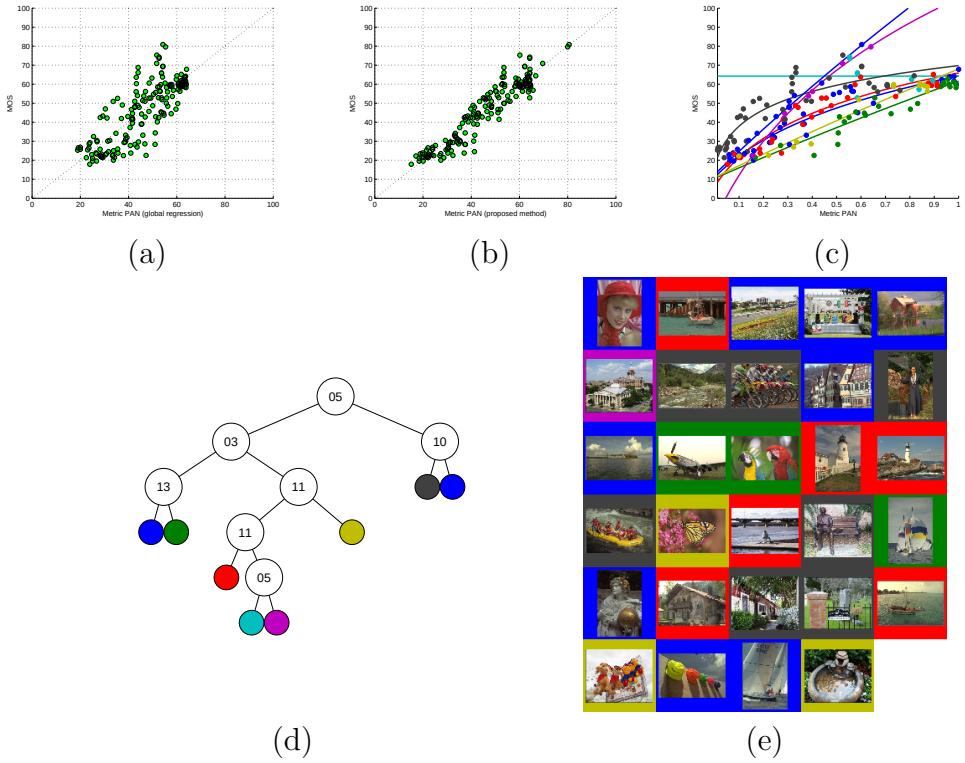


Figure 3.3: (a) scatter plot of the global regression between the PAN metric and the MOS scores; (b) scatter plot of the proposed method; (c) the cloud of points with the different instances of the model used; (d) the tree resulting from the pruning procedures; (e) the map of the distribution of the original images in the different leaves, represented by the background color. The colors have the same meaning in (c), (d) and (e).

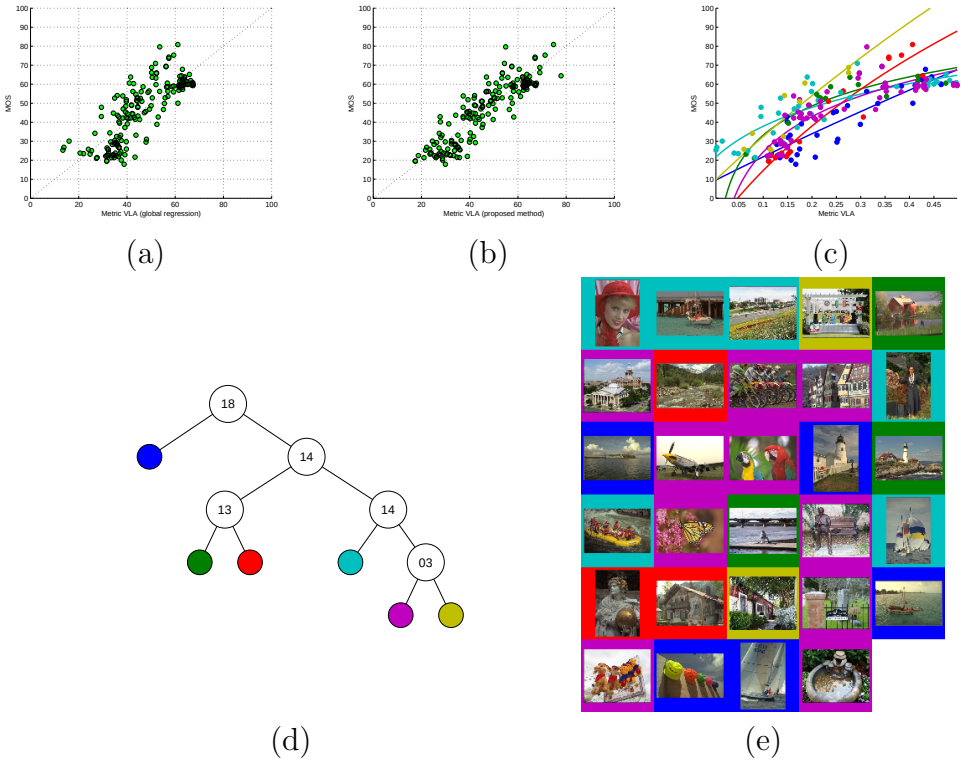


Figure 3.4: (a) scatter plot of the global regression between the VLA metric and the MOS scores; (b) scatter plot of the proposed method; (c) the cloud of points with the different instances of the model used; (d) the tree resulting from the pruning procedures; (e) the map of the distribution of the original images in the different leaves, represented by the background color. The colors have the same meaning in (c), (d) and (e).

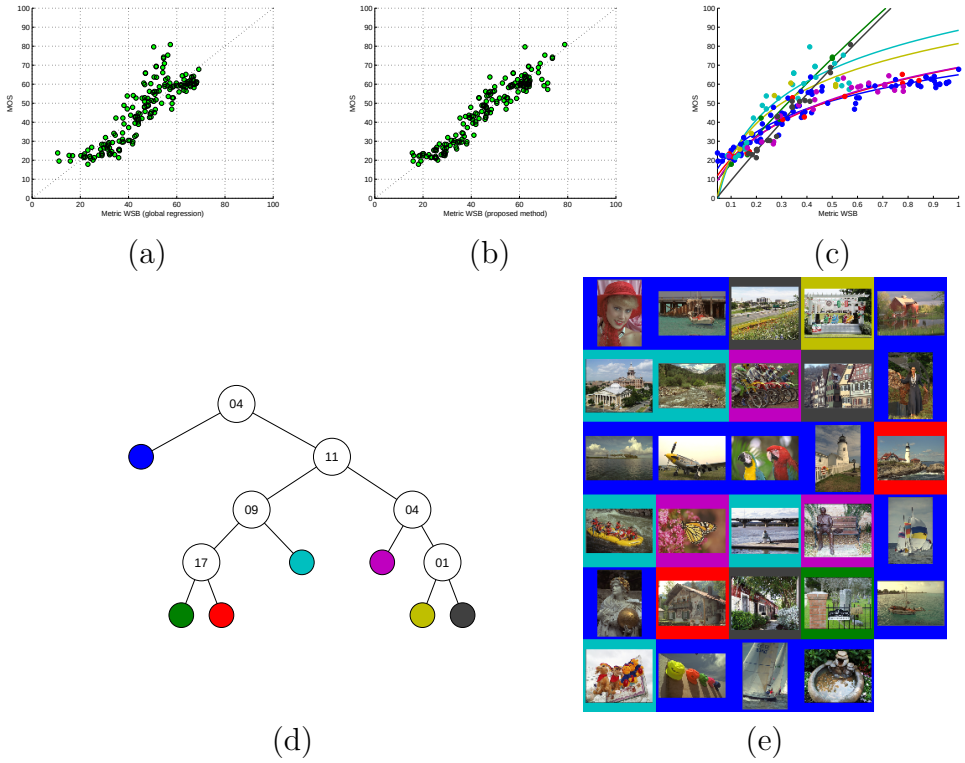
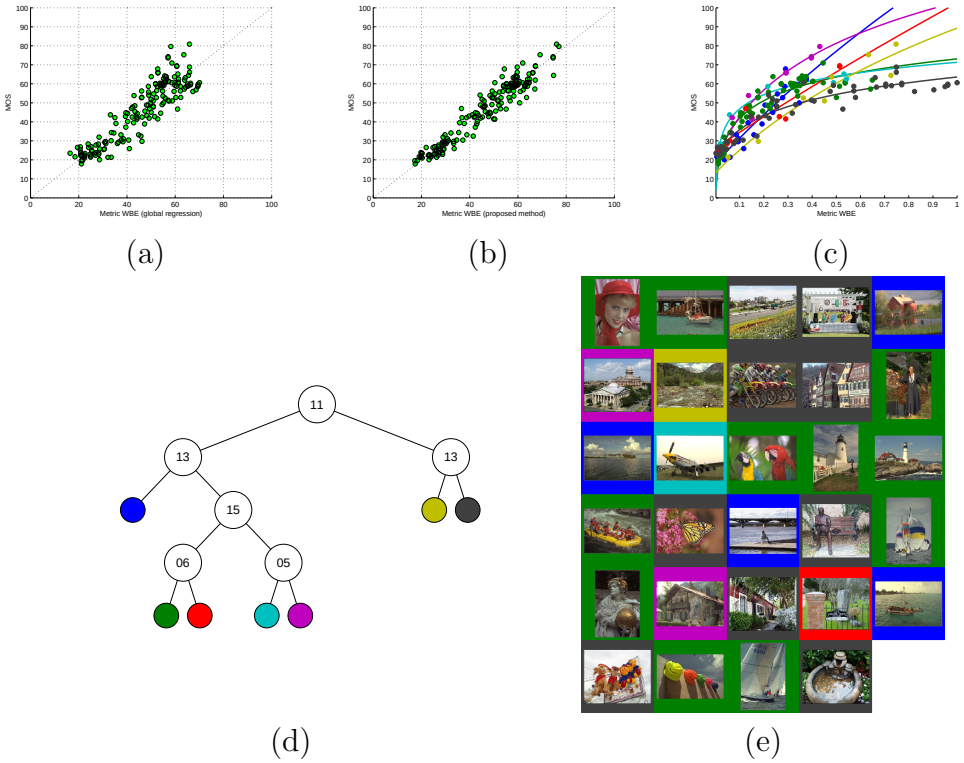


Figure 3.5: (a) scatter plot of the global regression between the WSB metric and the MOS scores; (b) scatter plot of the proposed method; (c) the cloud of points with the different instances of the model used; (d) the tree resulting from the pruning procedures; (e) the map of the distribution of the original images in the different leaves, represented by the background color. The colors have the same meaning in (c), (d) and (e).



(a) (b) (c)

(d) (e)

Figure 3.6: (a) scatter plot of the global regression between the WBE metric and the MOS scores; (b) scatter plot of the proposed method; (c) the cloud of points with the different instances of the model used; (d) the tree resulting from the pruning procedures; (e) the map of the distribution of the original images in the different leaves, represented by the background color. The colors have the same meaning in (c), (d) and (e).

Chapter 4

No-Reference Blur Metric

In this chapter we focus on No-Reference metrics for sharpness. Among the available methods found in the literature, after detecting the edge pixels, the sharpness measure is defined for each edge pixel. The final metric value is obtained averaging all these values [72, 6]. However, we have observed that in some cases this global measure is not representative of the real sharpness of the images. This fact is mainly due to the image noise that interferes with the measure at pixel-level. Pixel-level measures offer a poor signal-to-noise ratio that limit the accuracy of the local measurements. Performing the measure on a set of edge pixel can mitigate this problem. In the field of evaluation of digital imaging systems, the technique of slanted edge [2] cope the problem by integrating the measure along the edge of an properly designed pattern (Figure 4.1). In his PhD dissertation, Pham [83] proposed the extension of slanted edge measure to natural images by finding the straight line in the image through the use of an adaptive Hough transform. In our work [91] we further extend this approach by applying the measure to all the lines in the image (Figure 4.2). To implement our system we need to face the problem of identifying segments (groups of edge pixels) as support of our measures. In the Hough transform approach the property shared by edge pixels that belong to a certain segment is collinearity. In our system we need to define segments differently. Identifying a shared property between the edge pixels of a segment using a direct inspection of the edge map is problematic: while assuming that the pixels have to be spatially adjacent is straightforward, defining the starting point and the end point is difficult; so we choose a somehow complementary approach. In our work we segment the original image and extract all the boundaries between two different regions as

distinct segments. This solution guarantees the spatial adjacency property and produces segments bounded by two end points. Moreover, the pixels of the so defined boundaries, share the property of separating two coherent (with respect to the segmentation criteria) region of the image (Figure 4.3). In this chapter, we present an automatic method that allows to automatically select edge segments and permits to evaluate blurriness of the whole image on more reliable data.

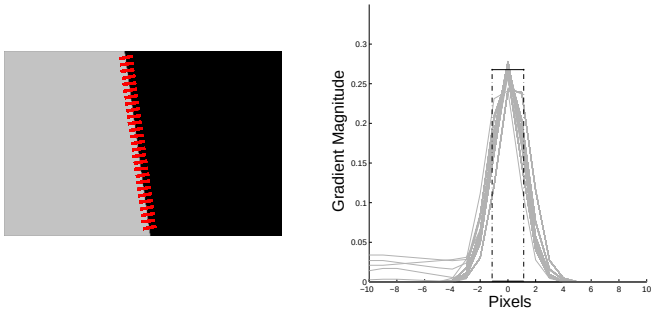


Figure 4.1: Slanted edge. All the profiles (red lines) extracted from the slanted edge pattern contribute to the estimate of the imaging system resolution.

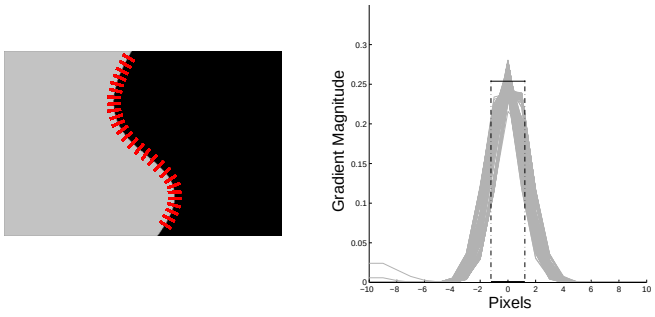


Figure 4.2: Our extension applied to a curved line.

Figure 4.3: A segmented image with one of the segment highlighted (red dotted line)

4.1 Mean Shift Segmentation

In our proposal, the automatic selection of the more representative edge segments starts from a region-based segmentation algorithm. Once these segments have been identified, an edge spread measure is defined to evaluate the image sharpness. An image segmentation is a partition of an image into contiguous regions of pixels that are similar in appearance. A large class of image segmentation algorithms is based on feature space analysis. In this paradigm the pixels are mapped into a color space and clustered, with each cluster delineating a homogeneous region in the image. Mean shift is a general nonparametric technique for the analysis of a complex multimodal feature space and the delineation of arbitrarily shaped clusters [24]. Mean Shift algorithm models a color image as a probability density function underlying a 5-dimensional space (three color channels and two lattice coordinates). The algorithm estimates the local density gradient using the offset of the mean vector computed in a window, from the center of that window. When the mean shift procedure is applied to every point in the feature space, the points of convergence aggregate in groups which can be merged. These are the detected modes (local maxima), and the associated data points define their basin of attraction. The clusters are delineated by the boundaries of the basins, and thus can have arbitrary shapes. The quality of segmentation is controlled by the spatial radius which determine the resolution in the 2D coordinate lattice domain, and the color radius which determine the resolution in 3D color space domain. The mean shift based color image seg-

mentation is already popular in the computer vision community and several implementations exist [17].

4.2 Profiles extraction

In our experiments we have used the Mean Shift algorithm to segment natural images. From the segmented image, we extract and collect all the boundaries between two adjacent regions as distinct segments. Given an edge segment of N edge pixels, we extract the N profiles along the direction of the gradient of each edge pixel, (see Figures 4.4a and 4.4c). We estimate the derivative of each profile using finite-differences, (see Figures 4.4b and 4.4d). After an alignment step (Section 4.3), we fit all the profiles with a Gaussian function (red line in Figures 4.4b and 4.4d). The standard deviation of this Gaussian is the blur estimation (spread) of the considered segment. The length of the profiles depends on the maximum edge spread we want to measure. In this work we have considered profiles of 21 pixels that permit a reliable estimation of sigma less than 5. This length was chosen with respect to the limited dimension of the images considered in our experiments (768 X 512 pixels).

To select the segment on which evaluating the overall blurriness of the image, we consider the following features:

- The length of the segment;
- The average contrast of the segment;
- The fitting error between the profiles and the Gaussian model.

In Figures 4.4a and 4.4c the selected segments are highlighted. Given a reliable segment, we have a redundant information about the edge spread over the N collected profiles. Therefore, we expect the estimation of this spread to be more stable with respect to noise.

4.3 Segment Spread Metric

All the profiles extracted with our method require a registration step before the fitting procedure (Figure 4.5). The registration is performed by identifying the zero-crossing point of the profile's derivative to identify the peak of the Gaussian. We then calculate the distance between the center of the

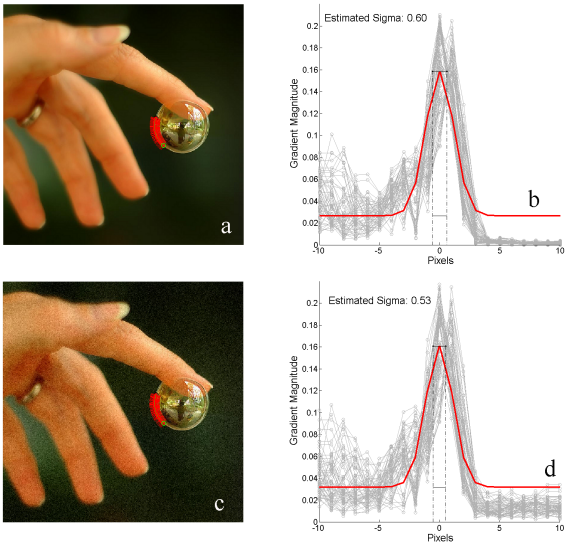


Figure 4.4: Our method applied to an images with and without noise

window and the zero-crossing point. Once the offset has been estimated, we translate the profile to the center of the window.

Our Segment Spread Metric (SSM) works as follows: We select the segments with contrast greater than 0.3, fitting error smaller than 0.03 and length greater than 30 pixels. We then extract the median of the sigmas of the selected segments. The threshold values 0.3,0.03 and 30 have been found empirically.

4.4 Results

We have performed our experiments on three datasets defined starting from the LIVE database [99]. The LIVE database contains a set of 145 images with different levels of blurriness. Our three datasets are composed as follows:

- $N0$ is the original LIVE dataset of 145 images;
- $N1$ consists of the 145 images of $N0$ plus the $N0$ dataset corrupted by a Gaussian noise with 16 gray levels of standard deviation (16 GLSTD) on the three channels, for a total of 290 images;

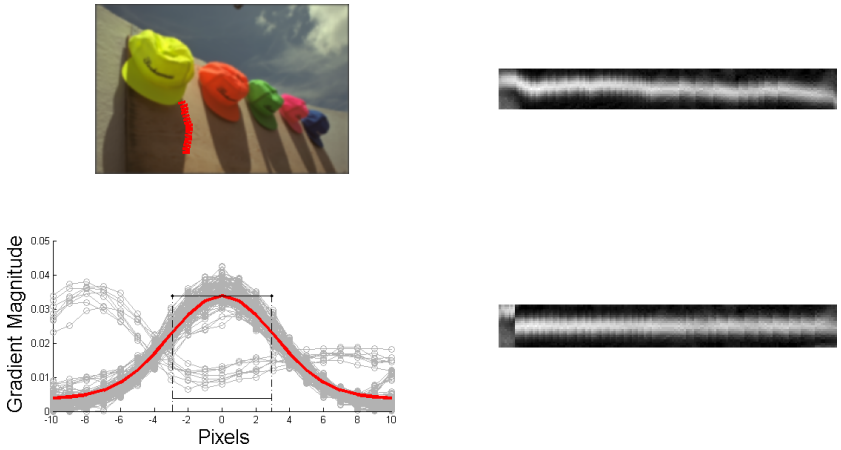


Figure 4.5: Top left, A segmented image with one of the segment highlighted. Top right, The extracted profiles. Bottom right, The extracted profiles after the registration procedure. Bottom Left, The plot of the registered profiles and the fitted model.

- $N2$ consists of the 290 images of $N1$ plus the $N0$ dataset corrupted by a Gaussian noise with 32 gray levels of standard deviation (32 GLSTD) on the three channels, for a total of 290+145=435 images.

We were interested in recovering from these images the sigma applied in the original LIVE database to generate the different level of blurriness. We have extended the LiVE sets introducing different levels of noise to test the stability of our SSM metric. SSM was compared with the Edge Spread Metric (ESM) proposed in [72]. In the presence of only blurriness among other available metrics, the ESM has proved to be more reliable metric [34]. Instead, ESM strongly suffers the presence of noise. Applying ESM to N1 and N2 datasets, we have obtained a constant value for all the images. This is due to its intrinsic procedure adopted to estimate the edge spread, which measures the distance between minimum and maximum values nearest to the edge point, along the gradient direction. High noise levels introduce false peaks which reduce this distance up to a constant value equal to one. Thus, to permit the metric comparison on the N1 and N2 datasets, we have modified the ESM introducing a pre-processing step, which consists in a convolution with a Gaussian filter. This filter was optimized for each of the N0, N1 and N2 datasets to reduce the MSE error between the ESM and the know applied sigma. Our method, instead, is not affected by the presence of noise and thus

Table 4.1: Regression parameters for SSM and EES. A: slope, B: Intercept

	SSM		ESM	
Name	A	B	A	B
N0	1.04	-0.10	0.29	-0.45
N1	1.10	-0.10	0.32	-0.77
N2	1.13	-0.03	0.34	-0.82

it does not need this pre-processing.

We have performed a first order polynomial regression on the estimated values of both the ESM and SSM. In Table 4.1 the slopes and intercepts obtained for the two metrics and for each of the three datasets are reported. Note that in the case of our measure the polynomials for all the datasets are the diagonal line.

As expected (section 4.3), our metric was unable to estimate the blurriness of all the images with blur greater than 7 (for a total of 21 images). While for images with sigma of 5.83 the metric does not provide a measure for 5 images on 6. In Table 4.2 we have reported the number of images were our metric was not able to provide a measure with respect to the level of blurriness and to the total of images with the same blurriness. From this Table we obtain that the total of images with no SSM response are 8, 17, 26 for the N0, N1 and N2 datasets respectively.

We have thus decided to remove all the images with no response to evaluate the performance of the ESM and SSM. In Figure 4.6 we have reported the scatter plots of the estimated sigmas, versus the applied sigmas for both ESM and SSM. Our predictions are less spread with respect to the diagonal line than those obtained with ESM. Finally the results obtained in terms of MSE between the estimated sigmas and the known applied blurriness are reported in Table 4.3 for both ESM and SSM and with respect to the three datasets. Our metric outperforms the ESM on all the datasets, as clearly indicated by the percentage of improvements.

4.4.1 Considerations on Depth of field

The overall image score can be influenced by the depth of field. Depth of field is the area in front of and behind the focus plane that is also “acceptably

Table 4.2: The number of images were our metric was not able to provide a measure, with respect to the level of blurriness and to the total of images with the same blurriness.

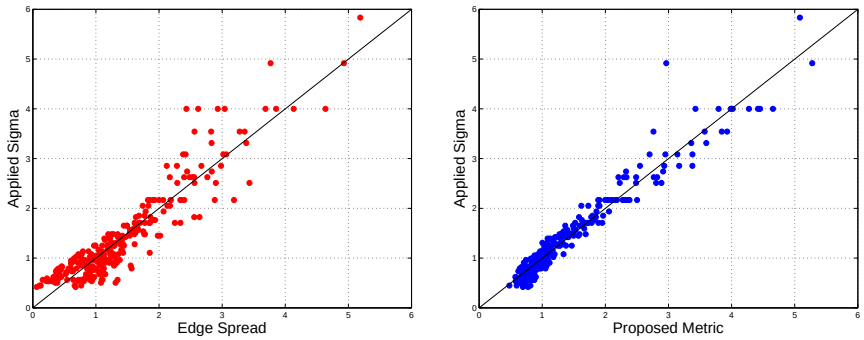
	No response		
Applied Sigma	N0	N1	N2
5.8333	1/2	3/4	5/6
7.6666	4/4	8/8	12/12
11.3333	2/2	4/4	6/6
15.0000	1/1	2/2	3/3

Table 4.3: Results: Gray level of standard deviation (glstd)

			MSE		
Name	Dataset	Cardinality	SSM	ESM	Improvements
N0	No noise	145	1.7	2.7	37%
N1	N0 + 16 glstd	145×2	3.6	5.7	36%
N2	N1 + 32 glstd	145×3	7.7	9.2	16%

sharp". [1] The extent of this area changes depending on the focal length, the focusing distance, and the aperture used. An example of image with low depth of field, where sharp edges and blurred edges are both present, is reported in Figure 4.7. In our test datasets there were few images with low depth of field; this explains why the median of the sigma is able to estimate correctly the applied blurriness. In a more general setup a better estimation could be obtained by using the minimum sigma.

[1] The depth of field change from sharp to unsharp as a gradual transition. Everything immediately in front of or in back of the focusing distance begins to lose sharpness, even if this is not perceived by the resolution of the camera. Since there is no a definite point of transition, a more rigorous term called the *circle of confusion* is used to define how much a point needs to be blurred in order to be perceived as blurred. When the circle of confusion becomes perceptible to the sensor, this region is said to be outside the depth of field and thus no longer "acceptably sharp".



(a) (b)

Figure 4.6: Scatter plots of the estimated sigmas, versus the applied sigmas for both ESM (a) and SSM (b)

Figure 4.7: The image is from imageCLEF database [75]. The median score 1.60 is not representative of the *perceived* sharpness.

Chapter 5

IQLab: Image Quality Assessment Tool

In this chapter we propose an image quality assessment tool. The tool is composed of different modules that implement several No-Reference metrics (i.e. where the original or ideal image is not available). Different types of image quality attributes can be taken into account by the No-Reference methods, like blurriness, graininess, blockiness, lack of contrast and lack of saturation or colorfulness among others. Our tool [90] aims to give a structured view of a collection of objective metrics that are available for the different distortions within an integrated framework. As each metric corresponds to a single module, our tool can be easily extended to include new metrics or to substitute some of them. The software permits to apply the metrics not only globally but also locally to different regions of interest of the image.

5.1 Tool Motivation

As cited by Sheik et al. [100] *All images are perfect, regardless of content, until distorted by acquisition, processing or reproduction.* In this way, we are implicitly assuming the presence of two signals: the content signal and the distortion signal. This philosophy assigns equal quality to all natural visual stimuli, and the task of NR Quality Assessment (QA) is reduced to blindly measuring the distortion.

Because a general model of the ideal image is not feasible, we have to

design a model for the different distortions. However, once we have properly designed such distortion model, the content can still influence the metric estimation.

For example, let us consider the image shown in figure 5.1a. The object in the foreground is visually sharped as desired by the photographer. The background, instead, is blurred on purpose (the camera settings were chosen to emphasize the depth of field). Probalbly, applying a NR metric to meaure the bluriness, a low quality score would be obtained while a subjective evaluation would give a higher score. This is because the metric is blind to the content of the image and can not distinguish between content signal and distortion signal. In this particular case, an ideal bluriness measure should be aware of the depth of field used in the photo.

a b

Figure 5.1: Two example images where a manual selection of the region of interest permits to reduce the content-distortion signal interference and makes the application of distortion metrics more reliable. Images are from the ImageClef database [75].

Psycho-visual experiments have shown that the perception of distortions is influenced by the amount of details in the images content [5]. This depen-

dency is coherent with the masking effects of the human visual system [80] specifically the texture masking effect.

If the goal is to obtain a reliable value when measuring a given distortion, we should validate the considered metric on a proper database that includes images representing all the possible contents that can interfere with the distortion to be measured. This is of course not feasible and the confidence of the result will diminish. Therefore, in order to obtain an objective score from the metric, this result should be interpreted as a function of the correlation between content and distortion in the particular image under study.

Instead of focusing on generating more representative databases, we choose an alternative way: we propose an interactive tool that permits the manual selection of the region of interest with respect to a given distortion. For example, if we have to measure the noise in figure 5.1b, the sky region should be selected; while for measuring the sharpness of the image, the object present in the scene (tower) should be the region of interest. To reduce the content-distortion signal interference, we propose here an interactive tool. The user can decide to apply a certain metric locally (the region of interest is manually selected) because the global one is not in correspondence with his subjective judgment. Moreover, he can also choose another NR method in case he was not satisfied with the previous result. This computer-aid process can be iteratively applied for each of the images.

Another purpose of our tool is to collect a dataset of images (and/or portions of them) with the corresponding numerical values of the different metrics considered. This dataset should be used in the validation process of these metrics to be correlated with psychophysical tests. In addition, this could give the hint to find a common scale to normalize the different metrics associated to a given distortion.

As above mentioned, images consist of the combining of two signals: content and distortion. As the distorsion increases, the visibility of the content decreases. We can thus represent images within a two dimensional space where the amount of "content" and "distortion" are taken into account. This image space is represented in figure 5.2a. High quality images (like for example those acquired by professional cameras, see figure 5.2b) occupy the left portion of this space. Their content is dominant with respect to the distortions. In the right portion of the space, we locate the images where the distortions are so significant that the content is recognized with difficulty (like for example figure 5.2d) and when applying a metric, we reasonably measure the distortion itself. With our tool we aim to evaluate the Quality

Assessment (QA) of the subset of images in the "intermediate range" where both content and distortion are significantly present and consequently, not easily decorrelated to be measured (like for example figure 5.2c). We note that the most of the natural images are located within this region.

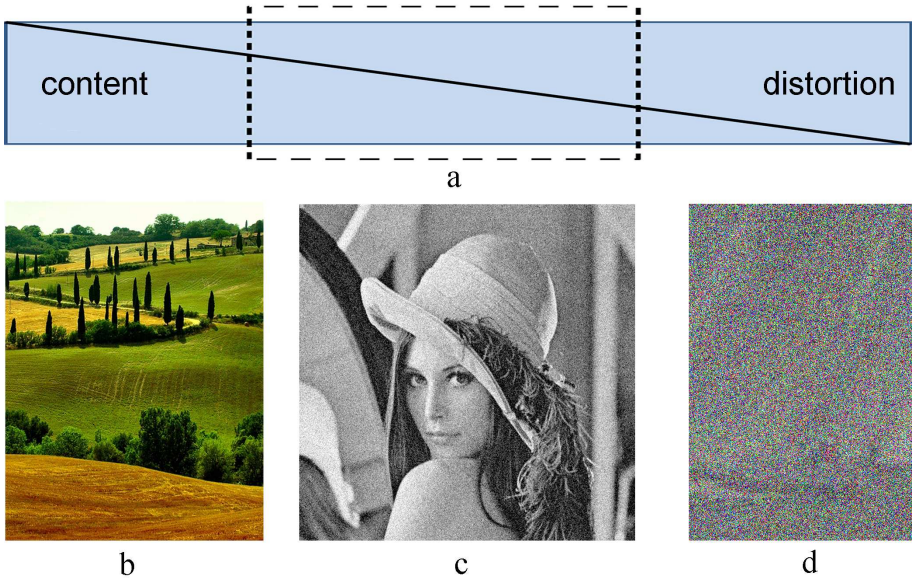


Figure 5.2: Images within the two dimensional space: content-distortion. High quality images ocupy the left portion of the space (b). In the opposite site, we find images where the distorsions are very high (d). The subset of images we address (c) belong to the region indicated within the dashed-line. Images (b) and (d) belong to the databases ImageClef [75] and LIVE [99] respectively.

5.2 Tool Description

For each browsed image, our tool permits to:

- apply each of the single metrics to the global image
- select a region of interest and apply locally each of the metrics
- visualize and collect in a table all the metrics applied to the image

- represent the different contributions of the artifacts, illustrating them in a pie chart

Up to now, the following metrics have been implemented within our tool:

- Bluriness:
  - Marziliano et al. [72] metric: it consists essentially of an edge detector. 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. Therefore, the edge width is measured and identified as the local blur measure. Finally, global blur is obtained by averaging the local blur values over all edge locations.
  - Crete et al. [26] metric: it is based on the discrimination between different levels of blur perceptible on the same image.
- Blocking artifacts:
  - Wang et al. [122] metric: it is defined in the frequency domain. The blocky image is modelled as a non-blocky image interfered with a pure blocky signal. The task of this algorithm is to detect and evaluate the power of the blocky signal.
  - Wang et al. [124] metric: it is a feature extraction method defined in the spatial domain. It measures the differences across block boundaries and zero-crossings.
  - Pan [79] method: it measures the horizontal and vertical inter-block differences. It takes into account the blocking artifacts for high bit rate images and the flatness for the very low bit rate images.
  - Vlachos [118] metric: designed in the frequency domain, where the blockiness measure is defined as the ratio between intra- and inter-block similarity.
- Noise:

  - Immerkaer [46] metric: the variance of additive zero mean Gaussian noise in an image is estimated. Different masks are considered : standard shifted differences, cascaded horizontal-vertical shifted differences, wavelet domain estimation, wavelet domain estimation with boundary removal, Immerkaer's method, Immerkaer's method with Daubechies-based Laplacian, Blockwise Immerkaer's method with Daubechies-based Laplacian.
  - Zhu and Milanfar [137] metric: it detects both blur and noise. The metric is based on the local gradients of the image and does not require any edge detection. Its value drops either when the test image becomes blurred or corrupted by random noise. It can be thought of as an indicator of the signal to noise ratio of the image.

- Global contrast:

  - Measure of enhancement EME [3]: it approximates an average contrast in the image by dividing the image into nonoverlapping blocks, defining a measure based on minimum and maximum intensity values in each block and averaging them.
  - Entropy: it indicates the occupation of intensity levels.

- White balance:

  - Gray world-based metric: The gray world algorithm [37] assumes that given an image of sufficiently varied colors, the average surface color in a scene is gray. This means that the shift from gray of the measured averages on the three channels corresponds to the color of the illuminant.
  - White Point-based metric: Assuming that there is always some white in the scene, the white point algorithm [59] looks for it in the image; its chromaticity will then be the chromaticity of the illuminant. The white point algorithm determines this white as the maximum R, maximum G and maximum B found in the image.

- Colorfulness:

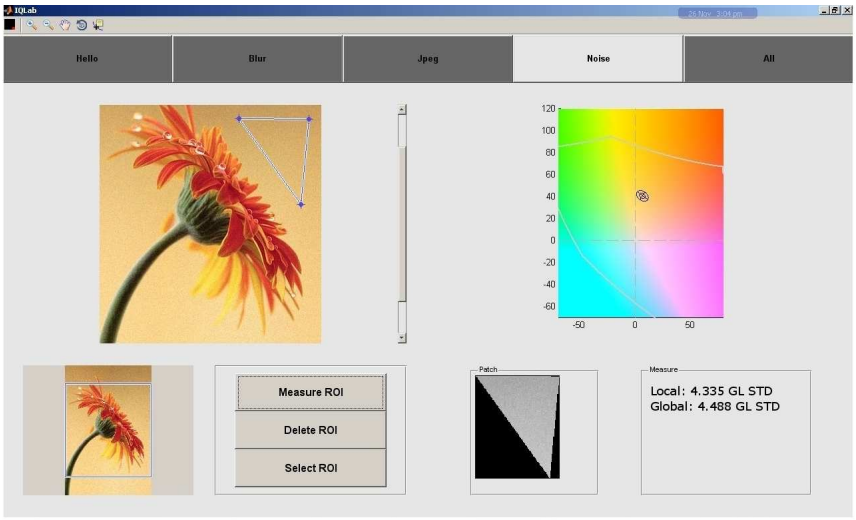
- Hasler and Susstrunk [42] metric: the distribution of the image pixels in the CIELab colour space is considered, assuming that the colourfulness can be represented by a linear combination of a subset of different quantities (standard deviation and mean of saturation and/or chroma). The parameters are found by maximising the correlation between experimental data and the metric.

As the tool is modular, other metrics can be easily added or used to substitute some of the above mentioned metrics.

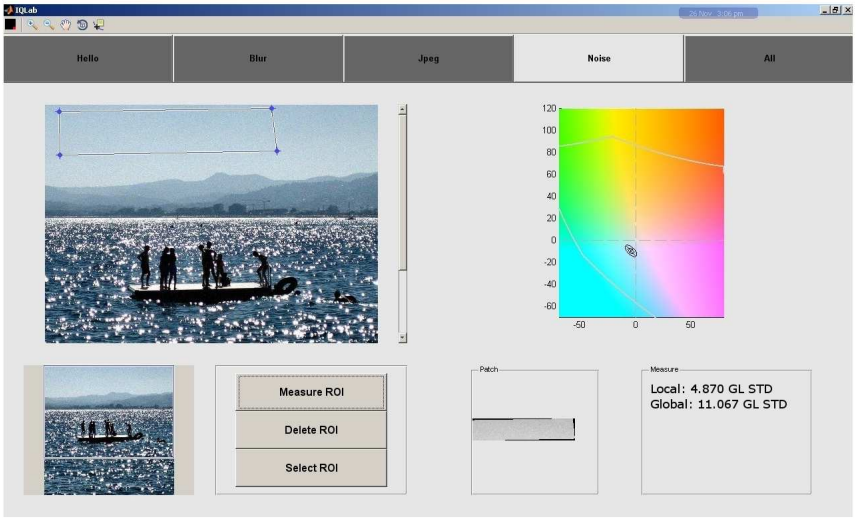
In what follows, two examples are reported. In the first example (see figure 5.3) we are interested in evaluating the noise. The global measure produces a value that could not be representative of the real noise of the image. For this reason, we permit the user to manually select the region where the metric has to be applied. The values obtained globally and locally could be significantly different. In figure 5.3 we apply the tool to two images where the same noise (equal to 4 Gray Level of Standard Deviation (GL STD)) was added. While the global values are significantly different (4.488 GL STD for figure 5.3a and 11.067 GL STD for figure 5.3b, noise measured with the method of Immerkaer [46] for the intensity channel), the local values are similar (4.335 and 4.870 GL STD). Therefore, the image quality of the two images could be better compared using the metrics locally.

The manually selected regions can be seen in detail in our interactive tool. The distribution of chromatic noise with respect to the ab-plane in the CIELab color space is also reported.

In the second case we are interested in measuring the sharpness. Again, after manually selecting the region of interest, the metric is locally applied. In the example of figure 5.4, the metric of Marziliano et al. [72] is applied. A new window permits to visualize the pixels considered to evaluate the edge spread within the selected region. On the main interface we can see how the edge spread is fit with a proper gaussian function.



a



b

Figure 5.3: Example of the tool interface. In this case we compare the results of the global and local metrics to evaluate the noise. The same synthetic noise (4 GL STD) is added to both original images (a and b). While the global values differ significantly, the local ones are similar and representative of the added noise. The regions of interest, manually selected, are highlighted.

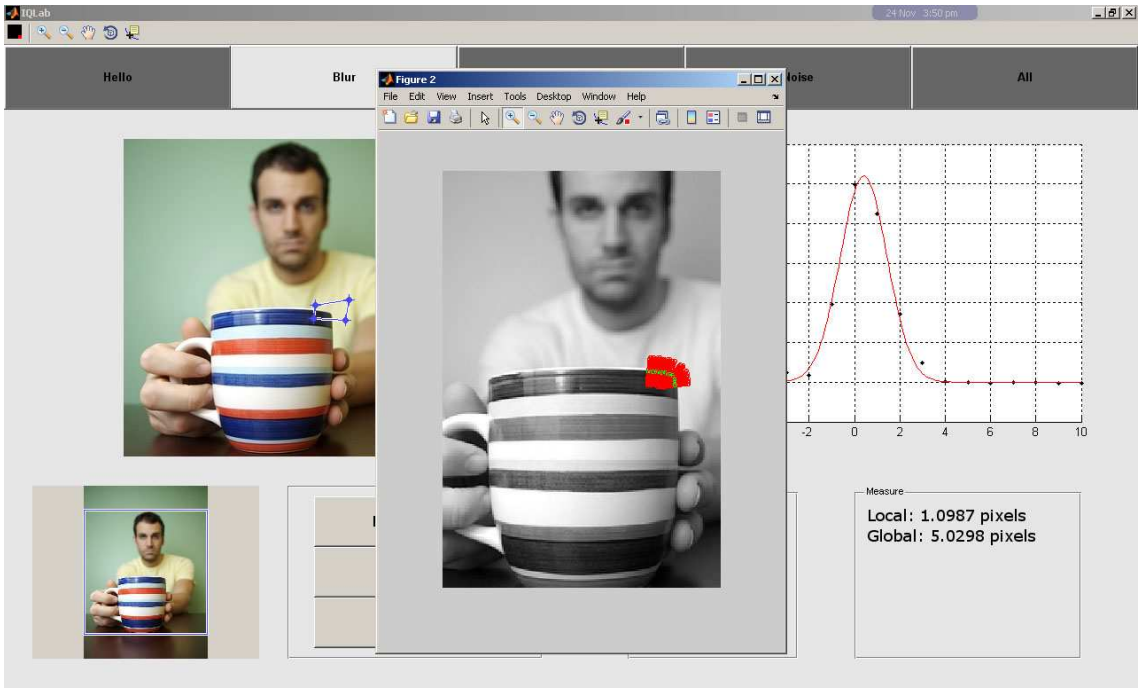


Figure 5.4: Example of the tool interface. In this case it is applied to evaluate the sharpness. The global and local edge spreads are reported. A new window (shown in foreground) shows the pixels used to calculate the edge spread. On the main interface (shown in background) the gaussian function used to approximate the edge spread is plotted.

Chapter 6

Conclusions

Objective visual quality assessment is a multi-dimensional research problem and an active and evolving research area. In Chapter 1, I have presented the state of the art of the different IQA methods and a guide for non-expert users in the choice and/or design of a workflow chain that has to make use of IQ metrics. In Chapter 2, I have addressed the problem of assessing the quality of demosaiced images. To this end I set up psycho-visual experiments to analyze the subjective evaluation of the artifacts introduced by the demosaicing process. The experiments were based on a dataset of distorted images generated starting from three color interpolation algorithms combined with two anti aliasing algorithms for a total of nine different methods. From the data analysis, it emerges that the perceptual quality of demosaiced images mainly depends on perceived sharpness, and on chromatic and achromatic zipper. The perception of the defects is more evident when the rendered images are compared with the reference one, while they may be unnoticed when images are evaluated alone. I have thus defined a new No-Reference metric for demosaicing artifacts based on measures of blurriness, chromatic and achromatic distortions that is able to fit these psycho-visual data. My metric can be applied to evaluate other demosaicing methods.

The perception of distortions is greatly influenced 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. This poses a problem when designing a No-Reference metric. Even if the metric was properly designed to identify a given artifact, the content of the image can still influence the metric estimation because the metric is blind to the pictorial

content of the image and thus can not distinguish between content signal and distortion signal. In Chapter 3, I proposed a new method to embed content information in image quality metrics. My assumption was that the content of the images alter the parameters of the regression model used to tune the objective metric. I tested this approach by extending four blocking metrics with image content information. My preliminary results show that this method outperforms the global regression obtained with the same model. Furthermore my approach suggests a separation between the design of the metric and its normalization/scaling with respect to the image content. . The relationship between model and metric should be further investigated to find some general properties required by the metric to be extended with my method. I strongly believe that this methodology can be extended in a straightforward manner to other kind of distortions.

A criteria to define the "best metric" for each distortion does not exist but this ideal metric should take into account the semantic and pictorial content of the image. Moreover the artifacts are intrinsically correlated as in the case of noise and blur. Trying to reduce the influence of one artifact could make the other more visible. This dependence requires designing Image Quality metrics that take into account different artifacts simultaneously. The great part of the Image Quality metrics is designed to measure only a single artifact while few consider two artifacts simultaneously. In Chapter 4, I have proposed a sharpness metric that measure image's blur by performing measurements on edge segments. Exploiting the information redundancy of these segment, my metric has shown to be robust against noise, and to be able to measure blur and noise independently of one another. This property can be used to face the problem of quality evaluation in case of multiple distortions. However, before considering different combination strategies, the normalization problem of the single metrics should be addressed. Therefore, both the normalization and combining of multiple metrics are still open problems within the Image Quality community.

Finally, the perception of distortions can be influenced by the local properties of the image signal. An Image Quality measure computed on the overall image may not be representative of the actual perceived quality. A more reliable way could be to compute the Image Quality measure on selected regions chosen on the basis of their properties and taking into account the application task. To investigate this aspect I have designed a modular No-Reference Image Quality Tool. The tool gives a structured view of a collection of objective metrics that are available for the different distortions

within an integrated framework. The Tool permits to apply the metrics not only globally but also locally to different regions of interest of the image. It also permits the collection of images and/or portions of them with the associated values of the different metrics. This collection may be used in psycho-visual experiments to correlate metrics and perceived distortions. Moreover, the tool permits to collect data useful to identify which of the different available metrics of a given distortion is more appropriate depending on the pictorial image content. For example, blur metrics that measure edge spread in the spatial domain are appropriate for isolated edges, while metrics in the frequency domain could be more suitable for the case of textured images. Another possible use of the above mentioned tool is to collect data to train a system that learns how to automatically select the region of interest for each distortion as we have shown in case blurriness. The choice of these regions of interest could be simplified applying a region annotation method.

Bibliography

[1] Methodology for the subjective assessment of the quality for television pictures. ITU-R Rec. BT. 500–11, 2002.

[2] International Standard ISO 12233:2000(E). *Photography Electronic still-picture cameras Resolution measurements*. ISO, Geneva, Switzerland.

[3] S. Agaian, K. Panetta, and Grigoryan A. Transform-based image enhancement algorithms with performance measure. *IEEE Trans. Image Process*, 10:367–382, 2001.

[4] Zhou Wang Alan, Zhou Wang, Alan C. Bovik, and Brian L. Evans. Blind measurement of blocking artifacts in images. In *in Proc. IEEE Int. Conf. Image Proc*, pages 981–984, 2000.

[5] E. Allen, S. Triantaphillidou, and Jacobson R. E. Image quality comparison between jpeg and jpeg2000. i. psychophysical investigation. *The Journal of imaging science and technology*, 51:548–258, 2007.

[6] Y. C. Chung J. M. Wang R. R. Bailey and S. W. Chen. A nonparametric blur measure based on edge analysis for image processing applications. In *Proc. IEEE Conference on Cybernetics and Intelligent Systems*, pages 356–360, 2004.

[7] B. Bayer. Color imaging array. U.S. patent 3971065, 1976.

[8] A. C. Bovik and Z. Wang. *Modern Image Quality Assessment*. Morgan & Claypool Publishers, 2006.

[9] A.C. Bovik and Shizhong Liu. Dct-domain blind measurement of blocking artifacts in dct-coded images. *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, 3:1725–1728, 2001.

[10] D. H. Brainard and D. Sherman. Reconstructing image from trichromatic samples: from basic research to practical applications. In *IS&T/SID Color Imaging Conference*, 1995.

[11] Toms Brando and Maria Paula Queluz. No-reference image quality assessment based on dct domain statistics. *Signal Processing*, 88(4):822 – 833, 2008.

[12] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Wadsworth, 1983.

[13] P.L. Callet and F. Autrusseau. *Subjective quality assessment IRCCyN/IOVC database*. http://www.irccyn.ec-nantes.fr/ivcdb/, 2005.

[14] M. Carnec, P. Le Callet, and D. Barba. Objective quality assessment of color images based on a generic perceptual reduced reference. *Signal Processing: Image Communication*, 23(4):239 – 256, 2008.

[15] D. Chandler and S. Hemami. A57 image database. http://foulard.ece.cornell.edu/dmc27/vsnr/vsnr.html, 2007.

[16] M. Choi, J. Jung, and J. Jeon. No reference image quality assessment using blur and noise. *International Journal of Computer Science and Engineering*, 2(3):76–80, 2009.

[17] Christopher M. Christoudias. Synergism in low level vision. In *Proceedings of the 16 th International Conference on Pattern Recognition (ICPR'02) Volume 4 - Volume 4*, ICPR '02, pages 40150–, Washington, DC, USA, 2002. IEEE Computer Society.

[18] A. Ciancio, A.L.N. da Costa, E.A.B. da Silva, A. Said, R. Samadani, and P. Obrador. Objective no-reference image blur metric based on local phase coherence. *Electronics Letters*, 45(23):1162 –1163, november 2009.

[19] G. Ciocca, C. Cusano, S. Santini, and Raimondo Schettini. Prosemantic features for content-based image retrieval. In *International Workshop on Advanced Multimedia Retrieval AMR 2009*, 2009.

[20] G. Ciocca, I. Gagliardi, and R. Schettini. Retrieving color images by content. In *Proceedings of the Image and Video Content-Based Retrieval Workshop*, pages 57–64, 1998.

[21] G. Ciocca, I. Gagliardi, and R. Schettini. Quicklook2: An integrated multimedia system. *Journal of Visual Languages and Computing*, 12(1):81 – 103, 2001.

[22] Erez Cohen and Yitzhak Yitzhaky. No-reference assessment of blur and noise impacts on image quality. *Signal, Image and Video Processing*, 4:289–302, 2010.

[23] D. R. Cok. Signal processing method and apparatus for producing interpolated chrominance values in a sampled color image signal. U.S. patent 4 642 678, 1986.

[24] D. Comaniciu and P. Meer. Mean shift analysis and applications. In *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, volume 2, pages 1197 –1203 vol.2, 1999.

[25] B. R. Corner, R. M. Narayanan, and S. E. Reichenbach. Noise estimation in remote sensing imagery using data masking. 24(4):689 – 702, 2003.

[26] F. Crete, T. Dolmiere, P. Ladret, and M. Nicolas. The blur effect: Perception and estimation with a new no-reference perceptual blur metric. In *Proc. SPIE Electronic Imaging Symposium Conf Human Vision and Electronic Imaging*, page 64920I, 2007.

[27] P.B. Crosby. *Quality is free*. McGraw-Hill, 1979.

[28] S. J. Daly. Visible differences predictor: an algorithm for the assessment of image fidelity. In B. E. Rogowitz, editor, *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, volume 1666 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, pages 2–15, aug 1992.

[29] H. de Ridder and S. Endrikhovski. Image quality is fun: Reflections on fidelity, usefulness and naturalness. *SID Symposium Digest of Technical Papers*, 33:986–989, 2002.

[30] M.P. Eckbert and A.P. Bradley. Perceptual quality metrics applied to still image compression. *Signal Processing*, 70(3):177–200, 1998.

[31] P. G. Engeldrum. Psychometric scaling:avoiding the pitfalls and hazards. In *IS&T's 2001 PICS Conference Proceedings*, pages 101–107, 2001.

[32] P.G. Engeldrum. A short image quality model taxonomy. *Journal of Imaging Science and Technology*, 48(2), 2004.

[33] U. Engelke, A. Maeder, and H.-J. Zepernick. Visual attention modelling for subjective image quality databases. In *Multimedia Signal Processing, 2009. MMSP '09. IEEE International Workshop on*, pages 1 –6, october 2009.

[34] R. Ferzli and L.J. Karam. A no-reference objective image sharpness metric based on just-noticeable blur and probability summation. In *Image Processing, 2007. ICIP 2007. IEEE International Conference on*, volume 3, pages III –445 –III –448, 16 2007-oct. 19 2007.

[35] Pina Marziliano Frederic, Frederic Dufaux, Stefan Winkler, Touradj Ebrahimi, and Genimedia Sa. A no-reference perceptual blur metric. In *IEEE 2002 International Conference on Image Processing*, pages 57–60, 2002.

[36] T. W. Freeman. Median filter for reconstructing missing color samples. *U.S. Patent 4724395*, 1998.

[37] G. and Buchsbaum. A spatial processor model for object colour perception. *Journal of the Franklin Institute*, 310(1):1 – 26, 1980.

[38] Salvador Gabarda and Gabriel Cristóbal. Blind image quality assessment through anisotropy. *J. Opt. Soc. Am. A*, 24(12):B42–B51, Dec 2007.

[39] Francesca Gasparini, Mirko Guarnera, Fabrizio Marini, and Raimondo Schettini. No-reference metrics for demosaicing. volume 7529, page 752911. SPIE, 2010.

[40] R. C. Gonzales and R.E. Woods. *Digital image processing*. Prentice Hall, 2008.

[41] M. Guarnera, G. Messina, V. Tommaselli, and A. Bruna. Directionally filter based demosaicing with integrated antialiasing. In *Consumer*

Electronics, 2008. ICCE 2008. Digest of Technical Papers. International Conference on, pages 1–2, 2008.

[42] D. Hasler and S. Susstrunk. Measuring colourfulness in natural images. In *Proc. SPIE Human Vision and Electronic Imaging*, pages 87–95, 2003.

[43] David Hasler and Sabine E. Süsstrunk. Measuring colorfulness in natural images. volume 5007, pages 87–95. SPIE, 2003.

[44] F. Idris and S. Panchanathan. Storage and retrieval of compressed images using wavelet vector quantization. *Journal of Visual Languages and Computing*, 8(3):289 – 301, 1997.

[45] Imatest. Digital image quality testing. `http://www.imatest.com/`.

[46] J. Immerkaer. Fast noise variance estimation. *Computer Vision and Image Understanding*, 64(2):300 – 302, 1996.

[47] John Immerkr. Fast noise variance estimation. *Computer Vision and Image Understanding*, 64(2):300 – 302, 1996.

[48] ISO. *Quality management and quality assurance. Vocabulary. ISO 84021994*. 2000.

[49] ITU. Methodology for the subjective assessment of the quality for television pictures. Technical report, ITU-R Rec. BT. 500-11, 2002.

[50] T. Janssen. *Computational Image Quality*. SPIE Press, 2001.

[51] T. Janssen and F. Blommaert. A computational approach to image quality. *Displays*, 21:129–142, 2000.

[52] Garrett M. Johnson and Mark D. Fairchild. Sharpness rules. In *Proc of IS&T/SID 8 th Color Imaging Conference*, pages 24–30, 2000.

[53] T. Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to predict where humans look. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 2106 –2113, 2009.

[54] J.M. Juran. *Juran on planning for quality*. The Free Press, Ney York, 1988.

[55] Gunturk B. K., Glotzbach J., Altunbasak Y., Schafer R. W., and Mersereau R. M. Demosaicking: Color filter array interpolation in single chip digital cameras. *IEEE Signal Processing Magazine (Special Issue on Color Image Processing)*, 2005.

[56] Vishwakumara Kayargadde and Jean-Bernard Martens. Perceptual characterization of images degraded by blur and noise: experiments. *J. Opt. Soc. Am. A*, 13(6):1166–1177, Jun 1996.

[57] R. Kimmel. Demosaicing: Image reconstruction from color ccd samples. *IEEE Transactions on Image Processing*, 8:548–258, 1999.

[58] T. Kusuma and H. Zepernick. On perceptual objective quality metrics for in-service picture quality monitoring. In *Third ATcrc Telecommunications and Networking Conference and Workshop*, 2003.

[59] E. H. Land. The retinex theory of color vision. *Scientific American*, 237(6):108–128, December 1977.

[60] V. Laparra, J. Muoz, and J. Malo. Divisive normalization image quality metric revisited. *J. Opt. Soc. Am. A*, 27(4):852–864, 2010.

[61] E.C. Larson, Cuong Vu, and D.M. Chandler. Can visual fixation patterns improve image fidelity assessment? In *Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on*, pages 2572 –2575, 2008.

[62] T. Leisti, J. Radun, T. Virtanen, R. Halonen, and G. Nyman. Subjective experience of image quality: Attributes, definitions and decision making of subjective image quality. In *Proc. of SPIE-IS&T Electronic Imaging*, volume 7242, 2009.

[63] Chaofeng Li and Alan C. Bovik. Content-partitioned structural similarity index for image quality assessment. *Signal Processing: Image Communication*, 25(7):517 – 526, 2010. Special Issue on Image and Video Quality Assessment.

[64] Qiang Li and Zhou Wang. General-purpose reduced-reference image quality assessment based on perceptually and statistically motivated image representation. In *Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on*, pages 1192 –1195, oct. 2008.

[65] X. Lia, B. Gunturk, and L. Zhang. Image demosaicing: A systematic survey. In *Proc. of SPIE*, volume 6822, 2008.

[66] P. Longere, Zhang Xuemei, P.B. Delahunt, and D.H. Brainard. Perceptual assessment of demosaicing algorithm performance. In *Proceedings of the IEEE*, volume 90, pages 123–132, 2002.

[67] W.M. Lu and Y.P. Tan. Color filter array demosaicing: New method and performance measures. *Image Processing IEEE*, 12:1194–1210, 2003.

[68] J. Lubin. A visual discrimination model for image system design and evaluation. In E. Peli, editor, *Visual Models for Target Detection and Recognition*, pages 207–220. World Scientific Publisher, 1995.

[69] Claes Lundstrom. Technical report: Measuring digital image quality. Technical report, Linkping UniversityLinkping University, Visual Information Technology and Applications (VITA), The Institute of Technology, 2006.

[70] Qi Ma and Liming Zhang. Saliency-based image quality assessment criterion. In De-Shuang Huang, Donald Wunsch, Daniel Levine, and Kang-Hyun Jo, editors, *Advanced Intelligent Computing Theories and Applications. With Aspects of Theoretical and Methodological Issues*, volume 5226 of *Lecture Notes in Computer Science*, pages 1124–1133. Springer Berlin / Heidelberg, 2008.

[71] Qi Ma, Liming Zhang, and Bin Wang. New strategy for image and video quality assessment. *J. Electronic Imaging*, 19(1):011019, 2010.

[72] Pina Marziliano, Frédéric Dufaux, Stefan Winkler, and Touradj Ebrahimi. Perceptual blur and ringing metrics: application to jpeg2000. *Sig. Proc.: Image Comm.*, 19(2):163–172, 2004.

[73] A. Mojsilovic, M.V. Popovic, and D.M. Rackov. On the selection of an optimal wavelet basis for texture characterization. *Image Processing, IEEE Transactions on*, 9(12):2043 – 2050, dec 2000.

[74] A.K. Moorthy and A.C. Bovik. Visual importance pooling for image quality assessment. *Selected Topics in Signal Processing, IEEE Journal of*, 3(2):193 –201, april 2009.

[75] Henning Müller, Paul Clough, Thomas Deselaers, and Barbara Caputo, editors. *ImageCLEF: Experimental Evaluation in Visual Information Retrieval*, volume 32 of *The Information Retrieval Series*. Springer, Berlin, 2010.

[76] A. Ninassi, O. Le Meur, P. Le Callet, and D. Barbba. Does where you gaze on an image affect your perception of quality? applying visual attention to image quality metric. In *Image Processing, 2007. ICIP 2007. IEEE International Conference on*, volume 2, pages II –169 –II –172, 2007.

[77] G. Nyman, J. Hkkinen, E. M. Koivisto, T. Leisti, P. Lindroos, O. Orenius, T. Virtanen, and T. Vuori. Evaluation of the visual performance of image processing pipes: information value of subjective image attributes. In *Proc. of SPIE-IS&T Electronic Imaging*, volume 7529, 2010.

[78] E. Ong, W. Lin, Z. Lu, X. Yang, S. Yao, F. Pan, L. Jiang, and F. Moschetti. A no-reference quality metric for measuring image blur. In *Signal Processing and Its Applications, 2003. Proceedings. Seventh International Symposium on*, volume 1, pages 469 – 472, july 2003.

[79] F. Pan, X. Lin, S. Rahardja, W. Lin, E. Ong, S. Yao, Z. Lu, and X. Yang. A locally-adaptive algorithm for measuring blocking artifacts in images and videos. In *Proc. of the 2004 International Symposium on Circuits and Systems*, volume 3, pages 925–928, 2004.

[80] T. N. Pappas, R. J. Safranek, and J. Chen. *Perceptual criteria for image quality evaluation*, chapter Handbook of Image and Video Processing, pages 939–959. Academic Press, 2005.

[81] Thrasyvoulos N. Pappas and Robert J. Safranek. Perceptual criteria for image quality evaluation. In *in Handbook of Image and Video Processing*, pages 669–684. Academic Press, 2000.

[82] Eli Peli. Contrast in complex images. *Journal of the Optical Society of America*, 7:2032–2040, 1990.

[83] T. Q. Pham. *Spatiotonal adaptivity in super-resolution of undersampled image sequences, Ph.D. thesis*. PhD thesis, Technische Universiteit Delft, Delft, The Netherlands, 2006.

[84] N. Ponomarenko, V. Lukin, A. Zelensky, K. Egiazarian, J. Astola, M. Carli, and F. Battisti. A database for evaluation of full reference visual quality assessment metrics. *Advances of Modern Radioelectronics*, 10:30–45, april 2009.

[85] J. Radun, T. Leisti, J. Hkkinen, H. Ojanen, J. Olives, and G. Nyman T. Vuori. Content and quality: Interpretation-based estimation of image quality. *ACM Transactions on Applied Perception*, 4, 2008.

[86] U. Rajashekar, I. van der Linde, A.C. Bovik, and L.K. Cormack. Gaffe: A gaze-attentive fixation finding engine. *Image Processing, IEEE Transactions on*, 17(4):564 –573, april 2008.

[87] K. Rank, M. Lendl, and R. Unbehauen. Estimation of image noise variance. *Vision, Image and Signal Processing, IEE Proceedings* -, 146(2):80 –84, aug 1999.

[88] F. Marini S. Corchs, F. Gasparini and R. Schettini. No-reference metrics for demosaicing. In *Proc. SPIE, Image Quality and System Performance VII*, volume 7529. SPIE, 2010.

[89] F. Marini S. Corchs, F. Gasparini and R. Schettini. No-reference metrics for jpeg: analysis and refinement using wavelets. In *Proc. SPIE, Image Quality and System Performance VII*, volume 7529. SPIE, 2010.

[90] F. Marini S. Corchs, F. Gasparini and R. Schettini. Image quality: a tool for no-reference assessment methods. In *Proc. SPIE*, pages 786712–786712–7,. SPIE, 2011.

[91] F. Marini S. Corchs, F. Gasparini and R. Schettini. A sharpness measure for automatically selected edge segments. In *Proc. SPIE, Image Quality and System Performance IX*, volume 8293. SPIE, 2012.

[92] N.G. Sadaka, L.J. Karam, R. Ferzli, and G.P. Abousleman. A no-reference perceptual image sharpness metric based on saliency-weighted foveal pooling. In *Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on*, pages 369 –372, 2008.

[93] R.J. Safranek and J.D. Johnston. A perceptually tuned sub-band image coder with image dependent quantization and post-quantization data compression. In *Acoustics, Speech, and Signal Processing, 1989.*

ICASSP-89., 1989 International Conference on, pages 1945 –1948 vol.3, May 1989.

[94] S. Saha and R. Vemuri. An analysis on the effect of image activity on lossy coding performance. In *Circuits and Systems, 2000. Proceedings. ISCAS 2000 Geneva. The 2000 IEEE International Symposium on*, volume 3, pages 295 –298 vol.3, 2000.

[95] Z.M.P. Sazzad, Y. Kawayoke, and Y. Horita. Mict image quality evaluation database. http://mict.eng.u-toyama.ac.jp/mict/index2.html, 2000.

[96] P. Scheunders, S. Livens, G. Van De Wouwer, P. Vautrot, and D. Van Dyck. Wavelet-based texture analysis. *Int. Journal of Computer Science and Information Management, Special issue on Image Processing (IJCSIM*, 1, 1997.

[97] G. Sharma. *Digital Color Imaging Handbook*. CRC Press, 2002.

[98] Gaurav Sharma. *Digital Color Imaging Handbook*. CRC Press, Inc., Boca Raton, FL, USA, 2002.

[99] H. Sheik, Z. Wang, L. Cormakc, and A. Bovik. Live image quality assessment database release 2. `http://live.ece.utexas.edu/research/quality`.

[100] H. R. Sheikh, A. C. Bovik, and L. Cormack. No-reference quality assessment using natural scene statistics: JPEG2000. *IEEE Transactions on Image Processing*, 14(11):1918–1927, 2005.

[101] H. R. Sheikh, M. F. Sabir, and A. C. Bovik. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Transactions on Image Processing*, 15(11):3440–3451, 2006.

[102] H.R. Sheikh and A.C. Bovik. Image information and visual quality. *Image Processing, IEEE Transactions on*, 15(2):430 –444, 2006.

[103] H.R. Sheikh, M.F. Sabir, and A.C. Bovik. A statistical evaluation of recent full reference image quality assessment algorithms. *Image Processing IEEE*, 15:3440–3451, 2006.

[104] H.R. Sheikh, Z.Wang, L. Cormack, and A.C. Bovik. Live image quality assessment database release 2. http://live.ece.utexas.edu/research/quality, 2005.

[105] Eero P Simoncelli and Bruno A Olshausen. Natural image statistics and neural representation. *Annual Review of Neuroscience*, 24(1):1193–1216, 2001.

[106] S. G. Smith. Color image restoration with anti-alias. *Patent US6842191*, 2005.

[107] Burkhard Stiller, Thomas Bocek, Fabio Hecht, Guilherme Machado, Peter Racz, and Martin Waldburger. Fundamentals and review of considered test methods. Technical report, I3A, 2007.

[108] Shan Suthaharan. No-reference visually significant blocking artifact metric for natural scene images. *Signal Processing*, 89(8):1647 – 1652, 2009.

[109] TASI. Technical advisory service for images. Technical report, 1979.

[110] Patrick Teo and David J. Heeger. Perceptual image distortion. In *in Proc. SPIE*, pages 982–986, 1994.

[111] L. L. Thurstone. A law of comparative judgement. *Psychological Review*, 34:273–286, 1927.

[112] Yubing Tong, Hubert Konik, Faouzi A. Cheikh, and Alain Tremeau. Full reference image quality assessment based on saliency map analysis. *Journal of Imaging Science and Technology*, 54(3):030503, 2010.

[113] W.S. Torgerson. *Theory and Methods of Scaling*. Wiley, Ney York, 1958.

[114] Antonio Torralba and Aude Oliva. Statistics of natural image categories. In *Network: Computation in Neural Systems*, pages 391–412, 2003.

[115] S. Triantaphillidou, E. Allen, and R. E. Jacobson. Image quality comparison between jpeg and jpeg2000. ii. scene dependency, scene analysis, and classification. *Journal of Imaging Science and Technology*, 51(3):259–270, 2007.

[116] Ian Van Der Linde, Umesh Rajashekar, C. Bovik, Alan, and Lawrence K. Cormack. Doves: a database of visual eye movements. In *Spatial Vision*, volume 22. http://live.ece.utexas.edu/research/doves, 2009.

[117] T. Vlachos. Detection of blocking artifacts in compressed video. *Electronics Letters*, 36(13):1106–1108, 2000.

[118] T. Vlachos. Detection of blocking artifacts in compressed video. *Electronics Letters*, 36(13):1106–1108, 2000.

[119] S.D. Voran and S. Wolf. The development and evaluation of an objective video quality assessment system that emulates human viewing panels. In *Broadcasting Convention, 1992. IBC., International*, pages 504 –508, July 1992.

[120] VQEG. Vqeg final report of fr-tv phase ii validation test. Technical report, Video Quality Experts Group (VQEG), 2003.

[121] C. T. Vu, E. C. Larson, and D. M. Chandler. Visual fixation patterns when judging image quality: Effects of distortion type, amount, and subject experience. *Image Analysis and Interpretation, IEEE Southwest Symposium on*, 0:73–76, 2008.

[122] Z. Wang, A. C. Bovik, and B. L. Evans. Blind measurement of blocking artifacts in images. In *Proc ICIP*, volume 3, pages 981–984, 2000.

[123] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, April 2004.

[124] Z. Wang, H. Sheikh, and A. C. Bovik. No-reference perceptual quality assessment of JPEG compressed images. In *Proc. IEEE International Conference on Image Processing*, pages 477–480, 2002.

[125] Zhou Wang, H.R. Sheikh, and A.C. Bovik. No-reference perceptual quality assessment of jpeg compressed images. In *Image Processing. 2002. Proceedings. 2002 International Conference on*, volume 1, pages I–477 – I–480 vol.1, 2002.

[126] Zhou Wang and Eero P. Simoncelli. Reduced-reference image quality assessment using a wavelet-domain natural image statistic model. In *in Proc. of SPIE Human Vision and Electronic Imaging*, volume 5666, pages 149–159, 2005.

[127] Zhou Wang, Eero P. Simoncelli, and Howard Hughes. Local phase coherence and the perception of blur. In *in Adv. Neural Information Processing Systems (NIPS03*, pages 786–792. MIT Press, 2004.

[128] A. B. Watson. DCT quantization matrices visually optimized for individual images. In J. P. Allebach & B. E. Rogowitz, editor, *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, volume 1913 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, pages 202–216, September 1993.

[129] Andrew B. Watson, Robert Borthwick, and Mathias Taylor. Image quality and entropy masking. In *SPIE Human Vision and Electronic Imaging Conference*, volume 3016, pages 2–12, 1997.

[130] S.R. Wayne. Quality control circle and company wide quality control. *Quality Progress*, pages 14–17, 1983.

[131] Arthur A. Webster, Coleen T. Jones, Margaret H. Pinson, Stephen D. Voran, and Stephen Wolf. An objective video quality assessment system based on human perception. In *in SPIE Human Vision, Visual Processing, and Digital Display IV*, pages 15–26, 1993.

[132] Chong-Yaw Wee, Raveendran Paramesran, R. Mukundan, and Xudong Jiang. Image quality assessment by discrete orthogonal moments. *Pattern Recognition*, 43(12):4055 – 4068, 2010.

[133] Stefan Winkler and Sabine Süsstrunk. Visibility of noise in natural images. In *Proc. IS&T/SPIE Electronic Imaging 2004: Human Vision and Electronic Imaging IX*, volume 5292, pages 121–129, 2004.

[134] Sergej Yendrikhovskij. Image quality: Between science and fiction. In *PICS*, pages 173–178, 1999.

[135] Junyong You, Andrew Perkis, Miska M. Hannuksela, and Moncef Gabbouj. Perceptual quality assessment based on visual attention analysis.

In *Proceedings of the seventeen ACM international conference on Multimedia*, MM '09, pages 561–564, New York, NY, USA, 2009. ACM.

[136] X. Zhang and B. A. Wandell. A spatial extension of cielab for digital color-image reproduction. *Journal of the Society for Information Display*, 5(1):61–63, 1997.

[137] X. Zhu and P. Milanfar. A no-reference sharpness metric senisitive to blur and noise. In *IEEE International Workshop on Quality of Multimedia Experience*, pages 64–69, 2009.