UNIVERSITA DEGLI STUDI DI MILANO-BICOCCA

Facoltá di Scienze Matematiche, Fisiche e Naturali
Dipartimento di Informatica, Sistemistica e Comunicazione

Dottorato di Ricerca in Informatica - XXV Ciclo

Image Quality Assessment for Digital Documents

Ph.D. Dissertation of: **Silvia Elena Corchs**

Supervisor: Prof. Raimondo Schettini
Tutor: Prof. Francesco Tisato
Ph.D. Coordinator: Prof.ssa Stefania Bandini

Anno Accademico 2013-2014

Contents

List of Figures

List of Tables

Acknowledgements

I want to express my sincere gratitude to my supervisor Prof. R. Schettini and all my colleagues of the Imaging and Vision Laboratory during the last four years: Fabrizio, Gianluigi, Simone, Alessandro, Claudio C., Claudio G., Paolo. A very special "thanks!" goes to Francesca.

The Doctoral study resulting in the present work was financially supported by Océ. The generation of the new database and the testing phase of the psychovisual experiments was done in collaboration with Océ Software Laboratories Namur. I wish to thank in particular Dr. M. Pracchi, B. Hucq and J. Bodart.

Abstract

This thesis focuses on No Reference (NR) methods for Image Quality Assessment (IQA). A review of the IQA field is presented in Chapter 2; where the different IQA methods are described and classified. In particular, the application of IQA methods within a workflow chain is discussed. In Chapter 3 we focus on NR metrics for JPEG-blockiness and noise artifacts. It is in general assumed that subjective methods produce an actual estimate of the perceived quality while objective methods produce values that should be correlated with human perceptions as best as possible. From the analysis of the regression curves that correlate objective and subjective data we have found that in some cases the metric's predictions are not in correspondence with the subjective scores. After reviewing the available databases, we realize that the distortion ranges considered are not in general representative of real case applications. Therefore, in Chapter 4 the Imaging and Vision Lab (IVL) database is introduced. It was generated with the aim of assessing the quality of images corrupted by JPEG and noise. In Chapter 5 we approach the NR-IQA field by focusing on a classification problem. A framework based on machine learning classification is proposed that let us evaluate how images can be classified within different groups or classes, according to their quality. NR metrics are considered as features and the assigned classes are obtained from the psychovisual data. For the JPEG distortion case, the feature space of the classifiers is built using each NR metric as single feature and also a pool of eleven NR metrics. Classification within five and three classes was addressed. In the former case, the five classes are in correspondence to the five categories recommended by the ITU (excellent, good, fair, poor, and bad) when designing image quality experiments. In the latter case we were interested in classifying images as high, medium or low quality ones. The classifiers are trained and tested on different databases. The classifier obtained using the pool of metrics outperforms each single metric classifier. Better performance is obtained in the case of three classes.

Considering an image as the combining of two signals, content and distortion, we note that the crosstalk between both signals influences both subjective and objective quality assessment. We address this problem in Chapter 6 where our working hypothesis is that regression can be improved if performed within a group of images that present similar contents in terms of low level features. The criteria chosen to divide the images in different groups is the image complexity. The proposed strategy consists on two steps: the images (of a given database) are first classified in three groups of low, medium and high complexity. In a second step, regression is performed within each of these groups separately. The strategy is tested for different NR metrics for JPEG-blockiness and noise artifacts, different databases are considered. Correlation coefficients are computed and statistical significance tests are applied. The gain in performance depends on the metric and distortion considered.

Summarizing, the two main proposals of this research work, i.e. the classification approach that combines several NR metrics and the grouping strategy, are able to outperform the correlation between subjective and objective data for the case of JPEG-blockiness. Both strategies can be extended to consider other type of distortions.

Chapter 1

Introduction

This thesis focuses on No Reference (NR) methods for Image Quality Assessment (IQA). In NR quality assessment, the algorithm does not have access to the reference image, and only the test image can be processed to assess its quality. The goal is to design methods whose evaluations are in close agreement with human judgments. From an application point of view, NR methods are more desirable than Full Reference (FR) methods (where the original image is available), which are mostly used for algorithms' testing and validation.

The most reliable way of measuring image quality is to ask human opinion. To this end, psychovisual experiments are conducted on distorted image databases where mean opinion scores are collected. The databases available in the literature have been created with the aim of validating IQ metrics.

The present research work originates with the collaboration with Océ Software Laboratories Namur in Belgium, who funded this PhD. The goal was to integrate IQA metrics within their products that focus on the preparation and printing of digital documents.

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

Given the variety of available objective metrics and databases, some questions arise: how do the metrics behave across different databases? Given a distortion type and several NR metrics, which of the metrics best measures the distortion? Does a combining of the NR metrics improve the single methods? In this thesis we investigate these issues, focusing mainly on JPEG-blockiness distortions.

In general, the distortion ranges of the available databases vary from images of high quality to images highly corrupted. However, in real applications it is not often to deal with so degraded images. Among the wide range of applications, NR IQA algorithms are certainly necessary for quality monitoring in real-time applications. For example, an IQA metric could be embedded within a printing workflow chain so that input images of high quality can be directly printed while low quality ones are discarded. Another example could be a web based image retrieval application where it could be helpful to automatically recover only images of high quality, in particular when dealing with huge databases. Having in mind these kind of applications, in the first part of this thesis we address the IQA field for high quality range and we approach it as a classification problem. To this end the Imaging and Vision Lab (IVL) database is introduced. The choice of the twenty reference images and the distortion range considered was done in collaboration with Océ researchers so as to represent

the images found in real cases as best as possible. The dataset generated is composed of 180 JPEG distorted images and 200 corrupted by white noise. Psychovisual experiments have been carried out where the assessment has been performed by observers belonging to Océ and from our Laboratory, i.e., all of them with image processing background. In Chapter 4 the Imaging and Vision Lab (IVL) database is introduced and described.

It is custom to apply non linear transformation to the metrics in order to better correlate with the subjective data. However, we find that some times the metrics' predictions are in disagreement with the subjective scores. Trying to improve the correlation we follow two different strategies. The first one is related to a classification task: can we obtain better performances classifying images within three or five quality groups instead of predicting precise quality scores? The second strategy consists in grouping the images in three classes according to their spatial complexity and then performing the regression analysis within each group separately.

In Chapter 5 we propose a framework based on machine learning methods to classify images within different groups or classes, according to their quality. NR metrics are considered as features and the assigned classes are obtained from the psychovisual data. Classification within five and three classes was addressed. In the former case, the five classes are in correspondence to the five categories recommended by the ITU [56] (excellent, good, fair, poor, and bad) when designing image quality experiments. In the latter case we were interested in classifying images as high, medium or low quality ones. The classifiers are trained and tested on different databases (LIVE [117], MICT [107] and IVL).

According to Sheik et al. [115]: *All images are perfect, regardless of content, until distorted by acquisition, processing or reproduction.* In this way, we are implicitly assuming that a digital image is the result of a combination of content and distortion signals. In high quality images (like for example those acquired by professional camera) the signal content is dominant with respect to the distortions. On the other hand, for low quality images the distortions are so significant that the content is recognized with difficulty and when applying a metric, we reasonably measure the distortion itself. In the intermediate range both content and distortion are significantly present and consequently, not easily decorrelated to be measured by IQA metrics. In this sense we can say that, in general, IQ metrics are not able to measure with the same performance the distortions within their possible full range and with respect to different image contents. Moreover, the crosstalk between content and distortion signals influences both the subjective and objective quality assessment. We address this problem in Chapter 6 where our working hypothesis is that the correlation between subjective and objective data can be improved if performed within a group of images that present similar contents in terms of low level features. The criteria chosen to divide the images in different groups is the image complexity. The proposed strategy consists on two steps: a first one consisting on a classification task, where the images (of a given database) are divided in three groups of low, medium and high complexity. In a second step, regression is performed within each of these groups separately. The strategy is tested for different NR metrics (specific for JPEG-blockiness and general purpose ones) and databases (LIVE, MICT and CSIQ [64]) and it is also applied to the case of Full Reference methods.

Summarizing, the thesis outline is as follows: the IQA field is reviewed in Chapter 2, in Chapter 3 we focus on NR metrics for JPEG-blockiness and noise artifacts, the new IVL database is presented in Chapter 4 where the psychovisual experiments performed are described. In Chapter 5 we present the machine learning based framework to classify images according to their quality while in Chapter 6 a complexity based grouping strategy is proposed to better correlate subjective and objective data. Finally the conclusions are drawn in Chapter 7.

Chapter 2

Image quality: state of the art

In this chapter a review of the Image Quality Assessment (IQA) field is presented. As it is done in recent review articles [18] the different IQA methods and algorithms are described and classified. The available IQA databases are briefly presented. Also the application of IQA methods within a workflow chain is discussed.

2.1 Image quality definitions

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 [58, 129, 155].

Image quality is often understood as the subjective impression of how well image content is rendered or reproduced [33]; the integrated set of perceptions of the overall degree of excellence of an image [35]; or an impression of its merits or excellence as perceived by an observer neither associated with the act of photographing nor closely involved with the subject matter depicted [61]. In these definitions image quality actually refers to the quality of the imaging systems used to acquire or render the images. Although suitable targets and studio scenes are often used for testing, we do not know in advance what objects/subjects will be actually acquired and processed. Depending on the applications, both scene contents and imaging conditions may range from being completely free (the common use of a consumer digital camera) or strictly controlled. 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"[52]; "fitness for (intended) use"[60]; "conformance to requirement"[28]; "user satisfaction"[148]. These definitions and their numerous variants could fit digital image quality 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."[123]. According to the International Imaging Industry Association[49], image quality 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 image quality has not been precisely defined, we can reasonably assume that in this case image quality requirements are low. On the contrary if the image were to "replace" the original, image quality requirements would be high. Taking into account that images are not be 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.

Janssen and Blommaert [59] state that in order to answer to the question what image quality is, it has to be split in other three questions: (1) what are images; (2) what are images used for; and (3) what are the requirements which the use of images imposes on them. To begin with the answers to the first two questions, they observe that images are the carriers of visual information about the outside world, and that they are used as input to human visual perception. Visual perception itself is part of the three processes perception, cognition, and action, which together constitute human interaction with the environment (see Figure 2.1). Images, therefore, can be regarded as input to the perception stage of interaction. Using a technical view of perception, it can be defined as the stage of human interaction which attributes of items outside the world are measured and internally quantified. The aim of this quantification is essentially two-fold. First, items in the outside world can be discriminated from one another using their internally quantified attributes. The result of this process is an essential step towards the construction of higher-level descriptions of scene geometry and object location, descriptions upon which later processes such as navigation in the scene are based. Second, items in the outside world can be identified by comparing their internally quantified attributes with quantified attributes, stored in memory, of similar items observed in the past. Identification of what is depicted in the image is an essential step in the interpretation of scene content and this determines the semantic awareness of what is in the scene. With respect to the third question, the authors conclude that the items depicted in the image should be successfully discriminable and identifiable. Summarizing, according to Janssen and Blommaert: 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 2.1 their schematic overview of the interaction process is shown. 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.

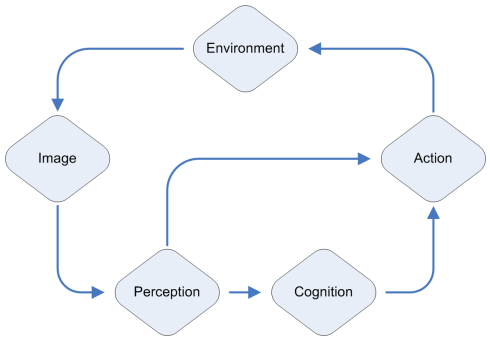

Figure 2.1: Schematic overview of the interaction process by Janssen and Blommaert [59].

2.2 Image Quality Modeling

Given a specific domain and task, there are several factors that may influence the perceived image quality:

- Factors intrinsic to the scene: scene geometry, lighting conditions, etc. . .

- Factors intrinsic to imaging devices: spatial resolution, geometric distortions, sharpness, noise, dynamic range, color accuracy, color gamut, etc...
- Factors depending by imaging processing pipelines: contrast, color balance, color saturation, compression, etc...
- Factors intrinsic to the human visual system: luminance sensitivity, contrast sensitivity, texture masking, etc...
- Factor depending by human observers: previous experiences, preferences and expectations, etc...

Different quality models have been proposed in the literature. For example, the Fidelity-Usefulness-Naturalness (FUN) IQ model [33] assumes the existence of three major dimensions: Fidelity, Usefulness and Naturalness.

Fidelity is the degree of apparent match of the image with the original (see Figure 2.2). 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 catalog require high fidelity of the images with respect to the originals. Genuineness and faithfulness are sometimes used as synonyms of Fidelity [49]. Dozens of books and hundreds of papers have been written about image fidelity and image reproduction e.g. [111].

Usefulness is the degree of apparent suitability of the 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 [42]. An example of image usefulness is shown in Figure 2.3. The image to the left may be accurate with respect to the original but the image to the right shows more details in the background due to a contrast enhancement algorithm applied. The enhancement processing steps have an obvious impact on Fidelity.

Naturalness is the degree of apparent match of the image with the viewer's internal references. This attribute plays a fundamental role when we have to evaluate the quality of an image without having access to the corresponding original. Examples of images requiring a high degree of naturalness are 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. Figure 2.4 shows three images with decreasing degrees of naturalness with respect to a mental reference of skin color.


a b c

Figure 2.2: Images exhibiting different fidelity degrees. a) Original image. b) Quantized image. c) Compressed image.

Recently, Moorthy et al. [80] suggested extending the dimensions of image quality by considering also its *Visual Aesthetic* and *Content*. We may refer to their model as the QAC model (Quality, Aesthetic, Content).

Figure 2.3: Example of image usefulness. a) A faithful image. b) A contrast enhanced image showing more details in the background.

Figure 2.4: Images with decreasing degrees of naturalness with respect to a mental reference of skin color

Visual aesthetics is a measure of the perceived beauty of a visual stimulus (see Figure 2.5). Aesthetics is intrinsically subjective, different users may consider an image to be aesthetically appealing for different motives based on their backgrounds and expectations. Notwithstanding the subjective nature of this dimension, several works tackle the problem to estimate the aesthetics of an image by developing computational procedures. These procedures exploit visual properties and compositional rules trying to predict aesthetic scores with high correlation with human perception [32, 102, 87].

Semantic content has an important impact on the evaluation of the quality of an image and it cannot be discounted during assessment (see Figure 2.6. Users' previous experiences influence the judgment of a good or bad image content. For example, an image can be considered of poor quality if it depicts offensive (for the user) content but if the same image is evaluated on the other quality dimensions it may receive a higher rating

It should be noted that, in general, the quality dimensions in the models are not independent. The overall IQ can be evaluated with metrics as a single number weighting the individual components. These weights depend on the specific image data type and on its function/goal.

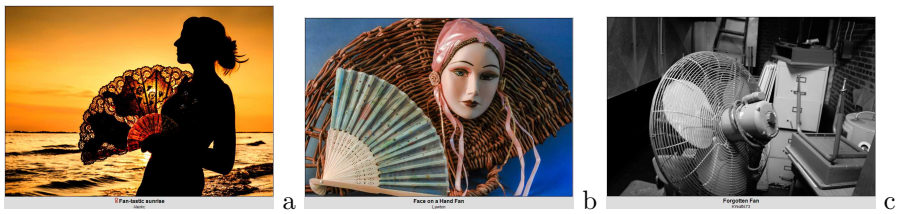

Figure 2.5: Examples of image aesthetic. The images are shown according to the aesthetic votes given by the community of the DPChallenge (http://www.dpchallenge.com) Web site. The subject refers to the "Fan" contest.

Figure 2.6: How image content influences quality. a) The image could be considered of poor quality because the tree was not fully captured. b) For a person hating spiders, the image may be not considered of good quality. c) An blurred image can be considered of good quality if the content is important for the photographer.

Summing up, the approach of creating an IQ model can be divided into three steps [8] :

- Identification of relevant quality attributes for the task at hand
- Determination of relationships between perceived quality values and objective measurements
- Combination of quality attribute measures to predict overall image quality

2.3 Image Quality Approaches

Image quality can be assessed either for an image seen in isolation or for an image seen together with a reference one. Image quality assessment is usually done by subjective and objective approaches.

2.3.1 Subjective approaches

The involvement of real people who view the images to assess their quality requires that all the factors that influence perception are taken into account to discount possible biases. To this end strict protocols have to be adopted. In the ITU standards [56], different subjective test methodologies are described. Regardless of the choice of the test methodology used, the way in which responses of the tests are analyzed depends upon the judgment (detection, etc.) and the information sought.

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

For both SC and SS methods there are different variants, the main difference is in the scale that the observers use to evaluate the presentations. For example, in adjectival categorical judgments, observers assign an image or image sequence to one of a set of categories that, typically, are defined in semantic terms. The categories may reflect judgments about the existence of a perceptible difference (e.g. “SAME”, “DIFFERENT”) or the existence and direction of perceptible differences (e.g. “LESS”, “SAME”, “MORE”). Categorical scales

that assess image quality and image impairment have been used most often, and in [56] readers can find some suggested scales to be used in the evaluation process. For each attribute/artifact, this method yields a distribution of judgments across scale categories.

In non-categorical judgments, observers assign a numerical value to each image or image sequence shown. These methods can have two kind of scales: continuous or discrete. In continuous scaling, a variant of the categorical method, the observer assigns each image or image sequence to a point on a line drawn between two semantic labels. The distance from an end of the scale is taken as the index for each presentation. In discrete scaling, the observer assigns each image or image sequence a number that reflects its judged level on a specified dimension (e.g. image sharpness). The range of the numbers used may be restricted (e.g. $0 - 100$) or not. Sometimes, the number assigned describes the judged level in absolute terms without direct reference to the level of any other image or image sequence as in some forms of magnitude estimation. In other cases, the number describes the judged level relative to that of a reference [61].

An important variant of the Stimulus Comparison is the Pairwise Comparison (PC) which is based on the law of comparative judgment studied by Thurstone [125]. In the PC method, the images are organized in sequences, each of which usually contains different versions of the same image. The images in each sequence are presented in pairs in two locations (for example one on the left and one on the right of the display) in all the possible combinations. Thus, each image is displayed twice in both locations. After each pair is presented, a judgement is made on which element in the pair is preferred based on some attributes. In PC, the tester doesn't impose any scale for the assessment. The selection of one image over the other is an exclusive Boolean choice.

The obtained scores can then be used as is, normalized using the mean and standard deviation to obtain Z-scores, or Thurstone scaling [125] can be used to create an interval scale, so that the scale represents equal perceptual distances. Finally the quality ratings from the evaluators are averaged to obtain the Mean Opinion Score (MOS) or the Difference Mean Opinion Score (DMOS). The latter is the difference between the MOS scoring of the test image and the MOS scoring of the corresponding reference image.

Let us note that it is also important to take into account the Human Vision System (HVS) characteristics, the image rendering procedure, the subjects characteristics and the perceptual task[35]. The HVS is specialized and tuned to recognize the features that are most important for human evolution and survival; there are other image features that humans cannot distinguish or that are easily overlooked[146]. These facts make quality assessment highly dependent on the image contents. Consider for example Figure 2.7. The same amount of Gaussian noise is applied to the image, first on the sky/clouds region (Figure 2.7a) and to the sand/rocks region (Figure 2.7b). The perceived image quality is strongly influenced by the distortion visibility. 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. This effect is called *Texture Masking* and is fundamental to take it into account when designing image quality metrics.

Subjective experiences and preferences may influence the human assessment of image quality; for example, it has been shown that the perceived distortions are dependent on how familiar the test person is with the observed image [42]. Image quality assessment is also affected by the user's task, e.g [49, 39]: 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 [36]. If the observer is provided with different instructions when evaluating a given image, he will give different scores to the same image


a


b

Figure 2.7: Example of how the perceptual quality is influenced by the visibility of the distortion. Gaussian noise is applied to the top (a) and bottom (b) regions of the image. The image in (b) is perceived as having higher quality than the image in (a).

depending on those instructions. Prior information regarding the image contents or fixation, may therefore affect the evaluation of the image quality.

Although effective, the efficiency of subjective approaches is very low. This has led the research towards the study of objective image quality measures not requiring human interaction.

2.3.2 Objective approaches

Objective approaches exploit suitable metrics computed directly from the digital image (see Figure 2.8). These image quality metrics can be broadly classified in Full Reference, No Reference, and Reduced Reference metrics [140].

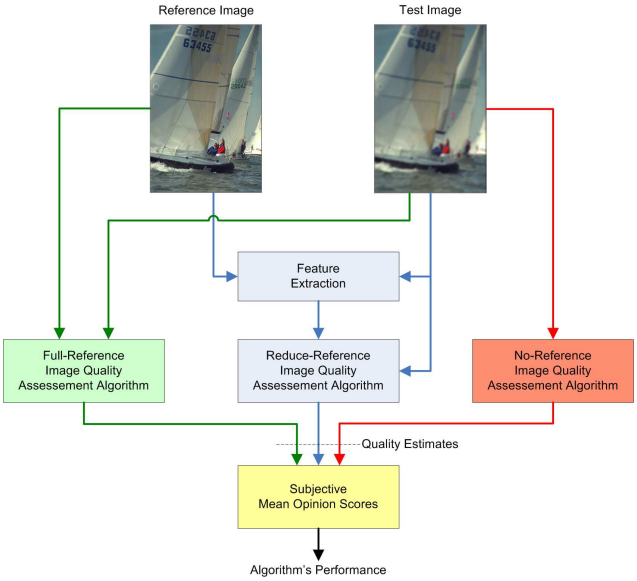

Figure 2.8: Objective image quality assessment approaches.

Full Reference Full Reference (FR) metrics perform a direct comparison between the image under test and a reference or "original" in a properly defined image space. 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 they are the most used, in general they do not correlate with subjective assessments [41, 139].

Error sensitivity frameworks follow a strategy of modifying MSE-like measures so that errors are penalized in accordance to their visibility. The evaluation of the visibility is accomplished by modeling some aspects of the HVS like Channel Decomposition, Contrast Sensitivity and Point Spread functions among others [30, 124, 72, 105]. All these techniques are bottom-up like approaches.

Top-down approaches take into account, for example, the image structure in defining the IQ since they assume that finding the structure is the goal for the cognitive process. The structural information in an image is defined as those attributes that represent the structure of objects in the scene, independently of the average luminance and contrast. The image quality is measured as a function of the amount of distortion that influence the image structure. Other approaches consider the characteristics of the natural images. They use natural scene statistics to quantify the loss of information due to the distortions present in the image.

A brief summary of FR metrics is presented in Table 2.1. The performance of each metric in terms of correlation with a ground truth, if available, is reported in the last column of the table (see Section 2.5).

No Reference No Reference (NR) metrics (also called blind methods) assume that image quality can be determined without a direct comparison between the original and the processed images.

Among the NR metrics we find those designed to identify the presence of specific processing distortions. Different types of defects can be considered: blurriness, the attenuation of the high spatial frequencies; blocking, discontinuities generated by block-based compression algorithms such as JPEG; graininess, random fluctuation of pixel values due to the device sensor; contrast, the difference in the brightness that makes an object in an image distinguishable from other objects and the background; colorfulness, the perceived difference between a color and gray. Blind methods can be classified as application-dependent since they are defined to handle with one or few specific defect types. Some of the blind methods are carried out in the frequency domain (like [24] for example) and make use of the common statistical characteristics of the power spectra of natural images [130] 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 Simoncelli and Olshausen [118]. More general-purpose NR IQA algorithms also exist which do not attempt to detect specific types of distortions. Methods of this type typically reformulate the IQA problem into a classification and regression problem in which the regressors/classifiers are trained using specific features. The relevant features are either discovered via machine learning or specified by using natural-scene statistics [81, 4, 78].

A brief summary of some NR methods is presented in Tables 2.2 and 2.3. When possible we report the overall performance score in the last column. If this score is not available, we report some performance scores of the most common defects either as a single value or as a range of values.

Reduced Reference Reduced Reference (RR) metrics lie between FR and NR metrics. They are designed to predict perceptual IQ with only partial information about the reference image. The methods extract a number of features from both the reference and the image

under test, and image comparison is based only on the correspondence of these features. Therefore, only image fidelity can be assessed. RR metrics may be useful to track the degree of visual degradation of image data that are transmitted through communication networks or during image acquisition. In image transmission the features must be coded and transmitted with the image data. The receiver compute the same features on the received image in order to verify if the original image has been corrupted during transmission. During image acquisition, in some image domains, it is common to acquire known targets (e.g. patches of colors or objects) on which compute the features to be evaluated. RR methods, in general, extract content-based or distortion-based features. Compared with FR and NR, few RR methods are available in the literature. In Table 2.4 a brief summary of RR methods is presented.

2.4 Available databases

Different standard databases are available to test the algorithms' performance with respect to the human subjective judgements. In what follows some of the most frequently used are briefly described:

- LIVE [117]: contains 29 reference images and 779 distorted images in 24-bpp color BMP format at different image resolutions ranging from 634 x 438 to 768 x 512 pixels.There are five distortion types in this database: JPEG compression (169 distorted images), JPEG2000 compression (175 distorted images), additive Gaussian white noise (145 distorted images), Gaussian blurring (145 distorted images), and JPEG2000 with bit errors via a simulatedRayleigh fading channel (145 distorted images). Each type of distortion was generated at 5-6 different amounts of distortion. The ratings were collected from 29 subjects. In order to visualize the distortion range, we show in 2.9 one original image from LIVE together with the most distorted versions for JPEG, blur and noise artifacts.

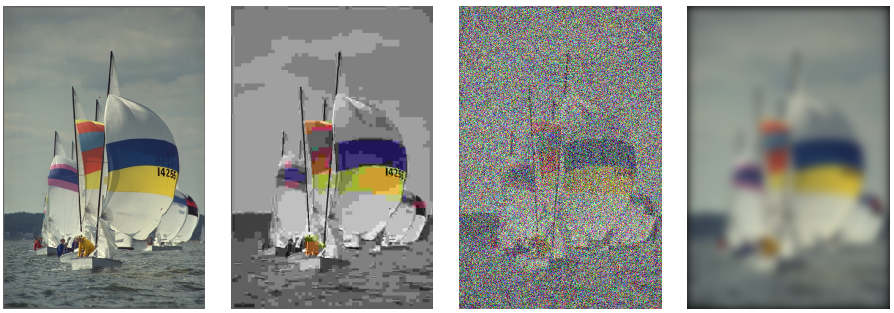

Figure 2.9: An original image from LIVE and its most distorted versions for JPEG, noise and blur artifacts respectively.

- MICT [107]: contains 14 reference images and 168 distorted images in 24-bpp color BMP format at a resolution of 768 x 512 pixels.There are two types of distortion in this database: JPEG compression (84 distorted images) and JPEG2000 compression (84 distorted images). Both types of distortion were generated at seven different amounts.

The ratings were obtained from 16 subjects. In order to visualize the distortion range, we show in Figure 2.10 one original image from MICT together with its most distorted versions for JPEG and JPEG2000 artifacts.

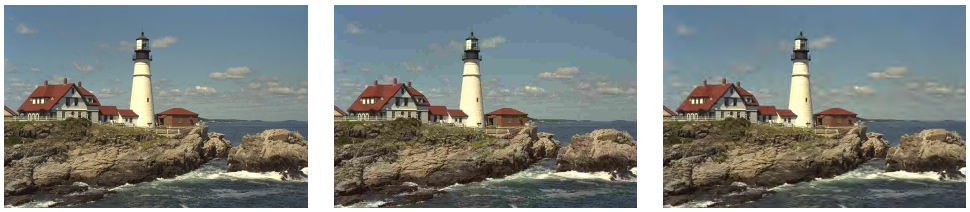

Figure 2.10: An original image from MICT and its most distorted versions for JPEG and JPEG2000 artifacts respectively.

- IRCCyN/IVC Scores on the MICT Database: Additional subjective ratings of the quality for the images from the MICT database were obtained by [131, 5] by using a different testing protocol, a different type of display, and different populations of subjects. The ratings were collected from 27 subjects.
- IVC [14]: contains 10 reference images and 185 distorted images in 24-bpp color BMP format at an image resolution of 512 x 512 pixels. There are five types of distortions in this database: JPEG compression (50 distorted images), JPEG compression of only the luminance component (25 distorted images), JPEG2000 compression (50 distorted images), locally adaptive-resolution coding (40 distorted images), and Gaussian blurring (20 distorted images). Each type of distortion was generated at five different amounts of distortion. The ratings were collected from 15 subjects.
- TID2008 [97]: contains 25 reference images and 1700 distorted versions. The reference images were obtained from the Kodak Lossless True Color Image Suite. All of the images are stored in 24-bpp BMP format at a resolution of 384 x 512 pixels. There are 17 distortion types in the database (e.g., different types of noise, blur, denoising, JPEG and JPEG2000 compression, transmission of JPEG, JPEG2000 images with errors, local distortions, luminance, and contrast changes). Each type of distortion was generated at four different amounts. The ratings were obtained from 838 subjects.
- A57 [17]: contains three original images and 54 distorted images (3 images x 6 distortion types x 3 contrasts). The distortion types considered are: additive Gaussian white noise, Baseline JPEG compression, JPEG-2000 compression using different settings, Gaussian blurring, quantization of the LH sub-bands of a 5-level DWT of the image.
- Categorical Subjective Image Quality (CSIQ) [64]: contains 30 reference images and 866 distorted images in 24-bpp PNG format at a resolution of 512 x 512 pixels. There are six distortion types in this database: JPEG compression (150 distorted images), JPEG2000 compression (150 distorted images), additive Gaussian white noise (150 distorted images), additive Gaussian pink noise (150 distorted images), Gaussian blurring (150 distorted images), and global contrast decrements (116 distorted images). Each type of distortion was generated at 4-5 different amounts. The ratings were obtained from 35 subjects. In order to visualize the distortion range, in Figure 2.11 one original image from CSIQ together with its most distorted versions for JPEG, JPEG200, blur, white noise, F noise and contrast artifacts are reported.

Figure 2.11: An original image from CSIQ and its most distorted versions for JPEG, JPEG200, blur, white noise, F noise and contrast artifacts respectively.

- LIVE Multiply Distorted [29]: it is the first datababase that addresses the multidistortion problem. It consists of 15 reference images and 405 multiply distorted images. Four levels of blur, JPEG compression and noise are considered. The multiple distorted images consist of blur followed by JPEG and blur followed by noise. The scores are collected from 37 observers.

For a more detailed description and review of these databases see [68] and [150].

2.5 IQA validation

Despite the time required to perform the test in a carefully controlled environment, subjective tests are at the base of objective quality metrics benchmarking. In fact, any objective

metric must be validated with respect to user judgements. The image quality databases serve as ground-truth information for evaluating IQA algorithms. Given a reference dataset, objective and subjective results can be compared through different performance measures. The Video Quality ExpertsGroup (VQEG) [136] recommends three performance criteria: prediction accuracy, prediction monotonicity and prediction consistency with respect to the subjective assessments:

- The prediction accuracy can be quantified either by measuring how well an algorithm predictions correlate with the subjective values or by measuring the average error between the algorithm predictions and the subjective scores. The Pearson Correlation Coefficient (PCC) and the Root Mean Squared Error (RMSE) are most commonly used for quantifying correlation and average error, respectively. For N data pairs (x_i, y_i), indicating with $\bar{x}$ and $\bar{y}$ the means of the respective data sets, the PCC is given by:

$$PCC = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}} \tag{2.1}$$

  Before computing PCC or RMSE, it is customary to apply a nonlinear transformation to the predicted scores so as to bring the predictions on the same scale as the subjective scores and to attempt to obtain a linear relationship between the predictions and opinion scores. The VQEG suggests the use of logistic or polynomial functions. The parameters of these functions are chosen to minimize the MSE between the set of subjective values (of a particular database) and the corresponding set of transformed predicted values. Recently, it has also been proposed a Monotonic Regression [44]. This function is obtained as an optimization problem that yields the highest PCC and does not depend on any parameter settings.

- The Spearman Rank Order Correlation Coefficient (SROCC) measures the prediction monotonicity of a metric, i.e. the degree to which the predictions of a metric agree with the relative magnitudes of the subjective ratings. Indicating with X_i and Y_i the ranks of x_i and y_i respectively, and with X' and Y' the midranks of the ordered data series, the SROCC is defined as:

$$SROCC = \frac{\sum(X_i - X')(Y_i - Y')}{\sqrt{\sum(X_i - X')^2}\sqrt{\sum(Y_i - Y')^2}} \tag{2.2}$$

- The Outlier Ratio (OR) is defined as the percentage of the number of predictions outside the range of ± 2 times the standard deviations of the subjective results. It measures the degree to which the metric maintains the prediction accuracy (i.e. prediction consistency). If N is the total number of data points and N' is the number of outliers, the OR is defined as:

$$OR = \frac{N'}{N} \tag{2.3}$$

Although the performance measures above described may give an idea of how well a given metric correlates with human perception, it may be misleading to only use them to select a metric to be used in a given domain for a given task. Considering for example the SROCC of different metrics on the same dataset, we can rank methods. However, as it can be seen from Tables 2.1-2.4 not all the metrics are validated on the same database. Tourancheau et al. [131] studied the impact of subjective dataset on the performance of IQ metrics. The authors

wondered if the objective metrics behaviors are constant across databases, contents and distortions and how significantly the subjective scores might fluctuate on different displays. To this end, the behaviors of four FR metrics (PSNR, SSIM, VIF and the metric by [15]) were tested on three image databases (LIVE, IVC, MICT). They demonstrated that the performances of the quality metrics can strongly fluctuate depending on the database used for testing and also showed the consistency of all metrics for two distinct displays.

2.6 Indirect quality evaluation

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

Image quality can also be indirectly assessed quantifying the performance of an image-based task performed by a domain expert and/or by a computational system. For example, in the framework of medical imaging, an image is of good quality if the resulting diagnosis is correct. In a biometrics system an image of a face may be considered of good quality is the person can be reliable recognized, in a Optical Character Recognition system a scanned document is a good quality if all the words can be correctly interpreted. The European Commission has proposed in 1999 an image quality standard for Computed Tomography images [34]. In this standard only two quality levels are considered: 1) Reproduction: Details of anatomical structures are visible but not necessarily clearly defined; and 2) Visually sharp reproduction: Anatomical details are clearly defined. Visual sharp reproduction does not affect the quality of the diagnosis. The quality evaluation could be done by processing each image and assessing the fulfillment of the constraints and requirements of the task [73]. This can be done manually by domain experts and/or automatically by a computational system. In the case of the face-based biometric system, the quality evaluation could be done by a face recognition algorithm that process and evaluates each image.

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

Indirect quality assessment can be carried out also by assessing the performance of the imaging/rendering devices. Using suitable sets of images and one or more direct methods (both objective and subjective), it is possible to assess the quality of the imaging and rendering procedures. In this case IQ is related to some measurable features of imaging/rendering devices, such as spatial resolution, color depth, etc. These features can be quantitatively assessed using standard targets (e.g. X-Rite ColorChecker® Classic [153], or the ISO 12233 Chart Data [54]) and ad-hoc designed software tools (e.g. [50]). However, these measures alone are not sufficient to fully assess IQ. The Camera Phone Image Quality (CPIQ) Initiative of the International Imaging Industry Association (I3A) uses both objective and subjective characterization procedures [49].

Figure 2.12 graphically depicts the different image quality approaches.

2.7 Applying IQA in a production workflow

After reviewing the different types and criteria used to classify the available metrics, we pose at this point the question of how to apply these metrics. Of course, the best general purpose

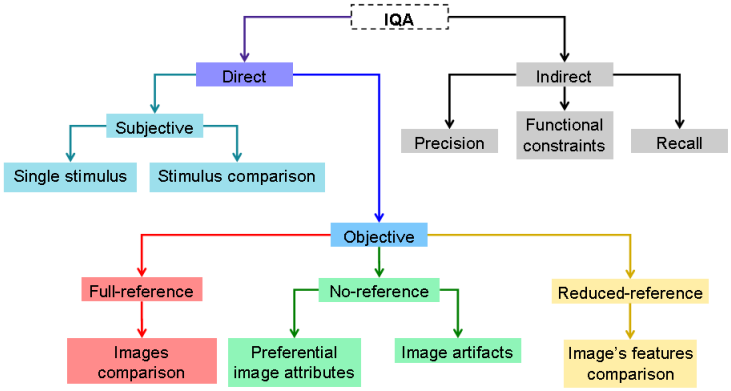

Figure 2.12: Taxonomy of the different image quality approaches.

metric does not exist. In general, different metrics may be required at different stages of the image production workflow chain. Even if a given task and scenario would require specific metrics, we can sketch some general guidelines.

In Figure 2.13 a generic image workflow chain with the indication of where the different IQA approaches are applied is shown. It starts with the source data (e.g. natural scene, phenomenon, measured values, etc.) to be captured by a digital image. The source can be specific of a narrow domain (e.g. resonance image) or broad domain (e.g. personal photo collections). Depending on the domain, acquisition constraints (e.g. semantic or environmental) are applied. The imaging block in Figure 2.13 broadly refers to any imaging device, hardware or software, that transform the source into a digital image. An example of a physical imaging device can be a camera, scanner or a tomograph while a software device can be any application that is able to create a synthetic image (e.g. map, flow chart, diagram, etc.). Once the image is created or acquired imaging metadata can be automatically embedded in the image header (e.g. EXIF). They may include information such as: maker, model of the camera, device settings, date and time, time zone offset, and GPS Information. Other metadata are usually added both for catalog and retrieval purposes. The overall metadata schema is usually set at the beginning of the digitalization stage and is based on application needs and the workflow requirements.

The digital image along with the metadata can be directly stored in an archive for further use or go through a validation phase that is aimed to have an initial assessment of the suitability and/or quality of the image with respect to the application needs. For example, a manual inspection can be performed in order to check if the whole scene has been correctly acquired or if it satisfies certain constraints. Validation constraints can be those related to the semantic of the image and can be evaluated either by human observers or automatically via computational algorithms (e.g. [110]). Images that do not pass the validation phase are rejected.

In the image quality 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 environ-

ment: geometry, lighting and surrounding. We may 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 device have an obvious impact on the quality of the acquired image. The hardware (e.g. sensors and optics) and/or 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 may 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 limit the images rejected in the validation phase. To this end RR or NR methods can be used to evaluate the device. Only RR methods can be used to evaluate the digital image with respect to the source because the source belongs to a different domain representation that makes it impossible a direct comparison. NR methods are used to detect the presence of defects in the imaging pipeline.

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. During this phase, any of the image quality techniques can be used. In particular, FR assessment techniques can be used since two digital images (before and after the processing) are available. Extra information can be added (usually information about the enhancements and processing that have been applied). The image can now be delivered and finally used either by a human observer or by an application.

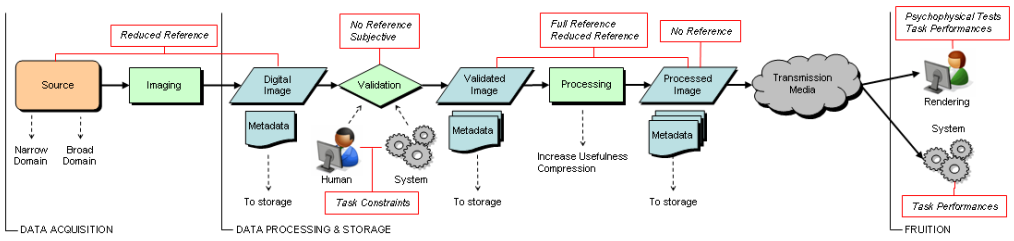

Figure 2.13: Relationship between the image production workflow chain and the image quality assessment approaches.

Given a processed image, we may be interested in predicting the overall IQ of the final printed document. The IQA of the image has to be evaluated before printing the document, so that the final product reaches the desired quality level. The processed image can be sent 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 judgments. 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).

During fruition the perceived image quality is greatly affected by the rendering device and the viewing conditions. For a faithful reproduction of digital images, the rendering devices must be carefully calibrated and characterized [112]. A best practice is to employ a color management system based on the International Color Consortium color management model. A device profile can be embedded into an image file which will enable the image to be automatically adjusted where necessary [53]. At this stage, image quality assessment can be carried out using subjective methodologies to evaluate the perceived image quality in the fruition environments. Some basic guidelines on how to acquire digital images of four different categories can be found in [10].

2.8 Conclusion and future research

Objective image quality assessment is an active and evolving research area. In the present chapter we have given a compendium of the state of the art of the different IQA methods. We have classified and summarized the different available metrics. We have also outlined the relationship between the image workflow chain and the image quality assessment approaches, how and when these different kinds of metrics can be applied within a generic image workflow chain. The selection and use of the different metrics depend on the semantic content of the image, the application task, and the particularly imaging chain applied. A challenge task in IQA is how to design a general purpose NR IQ metric capable of assessing different artifacts simultaneously. The major part of the NR IQ metrics are designed to measure only a single distortion. The few that consider two distortions simultaneously are mainly concerned with the case of noise and blur that are correlated. A possible strategy could be to combine different IQA metrics into a single method. However, before considering different combination strategies, the normalization problem of the single metrics should be addressed. Both the normalization and combination of multiple metrics are still open problems within the IQA community. The same issue applies if we aim to increase the performance of detecting a given artifact by combining several metrics. To cope with this problem, recently general purpose metrics (or universal metrics) have been proposed by [43], [4] and [154]. Despite these methods show promising results as generic metrics, they have been tested mainly on the LIVE database where each corrupted image is affected by a single distortion. Multi-distortion databases are thus required to further evaluate these novel metrics.

Finally, in order to design more reliable and general purpose image quality metrics, an interdisciplinary approach is the challenge for the next years. Evidence from the biological studies will help us to understand how our brain works when involved in the quality assessment task. Computational models of the visual system that account for these cognitive behaviors could be integrated within the perceptual quality metric design. Last but not least, semantic models coming from the image understanding community can certainly help us improve the metrics' design and performance.

Table 2.1: Full Reference Methods

FR Method	Brief description	Performance
MSE, PSNR	Measures the fidelity to the original and does not take into account HVS characteristics. It is the simplest and oldest measure. No parameters are needed.	–
Error sensitivity frameworks [30], [72], [105], [124], [147] (1989-1993)	Measure the fidelity to the original. These are bottom-up approaches that simulate functional properties of the HVS. Consist essentially in four modules: preprocessing (alignment, luminance transformation, and color transformation), channel decomposition (different choices are identity, wavelet, Discrete Cosine and Gabor transform), error weighting and error summation (Minkowski error pooling). Different parameters have to be estimated.	–
Spatial-CIELAB [159] (1997)	Measures color differences and is an 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.	–
Structure Similarity Index (SSIM) [140] (2004)	Measures the fidelity to the original. The HVS is adapted to extract structural information from natural visual scenes. Models image degradation as structural distortions instead of errors. The SSIM index is obtained as the product of three comparison components: luminance, contrast and correlation. Different parameters have to be estimated.	LIVE database. PCC = 0.967 RMSE = 5.06 OR = 0.041 SROCC = 0.963
Visual Information Fidelity Index (VIF) [113] (2006)	Measures the information shared between the two images. The construction of the VIF Index relies on the modeling of the statistical image source, the image distortion channel,and the human visual distortion channel. Different parameters have to be estimated.	LIVE database. PCC = 0.949. RMSE = 5.083. OR = 0.013. SROCC = 0.949.
Most Apparent Distortion Metric [64] (2010)	Combines two different strategies. For high quality images, local luminance and contrast masking are used to estimate detection based perceived distortion. On the other hand, changes in the local statistics of spatial-frequency components are used to estimate appearance-based perceived distortion in low-quality images.	TID, LIVE, MICT and CSIQ databases PCC = 0.8306 (TID), 0.9683 (LIVE), 0.8951 (MICT), 0.9502 (CSIQ) SROCC = 0.8340 (TID), 0.9675 (LIVE), 0.8908 (MICT), 0.9466 (CSIQ)
Divisive Normalization Metric [63] (2010)	Measures the closeness to the original. The metric is based on divisive normalization models [124] within Discrete Cosine Transform and Wavelet domains.	–
Discrete Orthogonal Moments [149] (2010)	Measures the Moment Correlation Index. Up to fourth order moments are computed on non-overlapping blocks for both the test and reference images. Correlation indexes are computed on each pair of block moments, and a single quality score is obtained by averaging all the correlation indexes. Two metrics are proposed: Q1 and Q2.	LIVE, A57, IVC and MICT databases. For Q1: PCC = 0.608-0.937 SROCC = 0.606-0.947 For Q2: PCC = 0.680-0.934 SROCC = 0.726-0.938
Machine learning approach [20] (2012)	It is based on a learned classification process in order to respect human observers. Support Vector Machine is applied for both classification and regression tasks. The feature vector contains visual attributes describing the images content.	LIVE and TID2008 databases. SROCC = 0.96 (LIVE), 0.90 (TID2008).

Table 2.2: No Reference Methods

NR Method	Artifacts	Brief description	Performance
[95] (1990)	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).	–
[51] (1996)	Noise	Estimates variance of the normally distributed noise.	–
[99] (1999)	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.	–
[135] (2000)	Blockiness	Designed in the frequency domain. The blockiness measure is defined as the ratio between intra- and inter-block similarity.	–
[141] (2000)	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.	–
[11] (2001)	Blockiness	Discrete Cosine Transform-domain algorithm. Blocking artifact modeled as a 2-D step function. Luminance and texture masking taken into account.	–
[142] (2002)	Blockiness	Feature extraction method in the spatial domain. Measures differences across block boundaries and zero-crossings. Non linear regression is applied where the parameters are estimated from subjective tests.	LIVE database. RMSE = 7.76. PCC =0.970. SROCC=0.960
[77] (2002)	Blur	Defined in the spatial domain. An edge detector is applied. For pixels corresponding to an edge location, the start and end positions of the edge are defined as the local extrema locations closest to the edge. The edge width is measured and identified as the local blur measure. Global blur obtained by averaging the local blur values over all edge locations.	105 images from LIVE and other sources. PCC = 0.85-0.96. SROCC = 0.87-0.96.
[26] (2003)	Noise	Laplacian and gradient data masks are used to estimate the additive and multiplicative noise standard deviations in an image. The histogram median value supplied the most accurate final noise estimations.	–
[46] (2003)	Colorfulness	Study of the distribution of the image pixels in the CIELab color 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 are found by maximising the correlation between experimental data and the metric.	84 images. PCC = 0.871-0.942.
[91] (2003)	Blur	The average edge spread in the image is measured by the average extent of the slope spread of an edge in both the gradients' direction and also the direction opposing the gradients' direction.	624 images. RMSE = 0.1774.
[94] (2004)	Blockiness	Measures horizontal and vertical inter-block difference. Takes into account the blocking artifacts for high bit rate images and the flatness for the very low bit rate images.	LIVE database. PCC = -0.930. SROCC = 0.932.
[145] (2004)	Blur	Defined in the frequency domain. Blur is interpreted as a disruption of the 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 behaviours of neurons in the primary visual cortex of mammals.	–
[151] (2004)	Noise	Investigates the visibility of noise itself as a target and uses natural images as the masker. Targets are Gaussian white noise and band-pass filtered noise of varying energy. Psychophysical experiments are conducted to determine the detection threshold of these noise targets on many different types of image content (noise visibility).	30 images. PCC = 0.95.

Table 2.3: No Reference Methods contd.

NR Method	Artifacts	Brief description	Performance
[37] (2007)	Blur and noise	The method is based on measuring the variance of the expected entropy of a given image on a set of predefined directions. Entropy can be calculated on a local basis 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.	–
[12] (2008)	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 [147].	LIVE database. RMS=7.439, PCC=0.973, SROCC=0.978
[22] (2009)	Blur and noise	Blur is estimated by the difference between the intensity of the current pixel and the average of neighbor pixels, the difference is normalized by the average.	LIVE database PCC = -0.91
[23] (2009)	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 [145], blur will introduce phase incoherence, causing these positions to change from sub-band to sub-band. 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.	6580 images with simulated and real blur. PCC = 0.5-0.75.
[121] (2009)	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 [137], the metric quantifies visual impairment by altering the spatial frequencies of the channels in order to standardize its sensitivity output so that it is independent from other channels.	Scores for 7 images in the LIVE Database. PCC = 083- 0.99.
[24] (2010)	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.	–
[69] (2011)	JPEG and JPEG2000	Neural Network-based approach. A feed-forward NN is employed to operate on the feature vector (blockiness and blur) extracted from JPEG/JPEG2000 images.	LIVE Database, JPEG: PCC=0.9623, RMSE=0.109 JPEG2000: PCC=0.930, RMSE=0.139
[43] (2011)	distortion generic	The method uses a set of low-level image features in a machine learning framework to learn a mapping from these features to subjective image quality scores. Features are derived from natural image statistics, texture features and blur/noise estimation.	LIVE Database, PCC=0.89.
[38] (2012)	Gaussian noise and Gaussian blur	The von Mises distribution of the image information is evaluated. Assuming that the concentration parameter decreases exponentially with increasing the amount of degradation, it can be used as an image quality assessment index.	TID2008 Database, Noise: PCC=0.8052, SROCC=0.8083 Blur: PCC=0.9600, SROCC=1
[4] (2012)	distortion generic	The Blind/Referenceless Image Spatial QUality Evaluator (BRISQUE) operates in the spatial domain and is based on natural scene statistics. The algorithm quantifies the naturalness in the image due to presence of different distortions.	LIVE Database, SROCC = 0.9395, PCC = 0.9424.
[154] (2012)	distortion generic	Approach based on visual codebooks. A visual codebook consisting of Gabor-filter-based local features extracted from local image patches is used to capture complex statistics of a natural image. The codebook encodes statistics by quantizing the feature space and accumulating histograms of patch appearances.	LIVE Database, PCC = 0.0215, SROCC = 0.0199.

Table 2.4: Reduced Reference Methods

RR Method	Features	Brief description	Performance
[106] (2000)	Features describing aliasing and blockiness effects.	The active regions of an image (defined as those with strong edges and textures) are quantified. The metric is based on the wavelet coefficients from the different sub-band coding schemes and is used to predict the PSNR of compressed images.	R-Squared value = 0.9934
[62] (2003)	Features describing blocking and blurring artifacts.	Hybrid IQ metric. The importance of blocking effect is computed using the Wang and Bovik method [141], and the importance of blurring is measured using Marziliano's method [77].	–
[143] (2005)	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.	LIVE database. PCC = 0.9695 (JPG), 0.8889 (noise), 0.8872 (blur), 0.9353 (JPG2K). SROCC = 0.8908 (JPG), 0.8639 (noise), 0.9145 (blur), 0.9298 (JPG2K). OR = 0.0341 (JPG), 0.1793 (noise), 0.1172 (blur), 0.069 (JPG2K).
[15] (2008)	Visual features similar to those used by the HVS: orientation, length, width and magnitude of the contrast at the characteristic point.	Implements an operating and organizational model of the HVS, including important stages of vision (perceptual color space, CSF, psychophysical sub-band decomposition, masking effect modeling). The criterion extracts structural information from the representation of images in a perceptual space. Extracted features are stored in a reduced description which is generic, as it is not designed for specific types of distortions.	IVC, LIVE and MICT databases. PCC = 0.913-0.972. ROCC = 0.909-0.953. OR = 0.02-0.05.
[67] (2008)	Statistical features extracted from a divisive normalization-based image representation.	Inspired by the success of the divisive normalization transform as a perceptually and statistically motivated image representation. Each coefficient of the transform is normalized (divided) by the energy of a cluster of neighboring coefficients. It is a general-purpose method, no assumption is made about the types of distortions present in the images.	LIVE database. PCC = 0.9162 SROCC = 0.9279 OR = 0.1079
[119] (2012)	Entropy of Wavelet coefficients	The algorithm measures the changes in suitably weighted entropies between the reference and distorted images in the wavelet domain.	LIVE and TID2008 databases. SROCC = 0.8606 (LIVE) 0.824 (TID2008)

Chapter 3

No Reference metrics for JPEG-blockiness and noise distortions

3.1 NR metrics for JPEG-blockiness

One of the image quality distortions for which several objective metrics have been developed is blockiness. The blocking artifact is a prevailing degradation caused by the block-based Discrete Cosine Transform coding technique, especially under low bit-rate conditions, due to the different quantization sizes used in the neighboring blocks and the lack of consideration for inter-block correlation.

Numerous NR IQA algorithms have been developed specifically for JPEG images. The general approach involves measuring the edge strength at block boundaries and then using this measure to estimate the visibility of the blocking, often based on masking. Quality is then determined based on this estimate of perceived blockiness.

In what follows we list some of the frequently used JPEG-blockiness measures.

- **M1**, Generalized Block-edge Impairment (GBIM) metric developed by Wu and Yuen [152]: It is the most well known metric in the spatial domain. If an image I of size W x H is divided into B x B blocks, the horizontal and vertical difference (discontinuity) at block boundaries can be evaluated as:

$$M_h = \left[\sum_{k=1}^{H/B-1} \sum_{x=0}^{W-1} (I(x, k.B-1) - I(x, k.B))^2 \right]^{1/2} \tag{3.1}$$

for horizontal blockiness, and

$$M_v = \left[\sum_{l=1}^{W/B-1} \sum_{y=0}^{H-1} (I(l.B-1, y) - I(l.B, y))^2 \right]^{1/2} \tag{3.2}$$

for vertical blockiness. In general it is assumed that B=8 because this is the most frequently used block size for block transform based image coding. The two directions

are then combined into a single quality value. The higher the GBIM value is above one, the greater the severity of the blocking effect. Variations of the GBIM method can be found in [79, 92]. Object edges at block boundaries can be excluded in blockiness consideration [158].

- **M2**, developed by Vlachos [135]: the algorithm uses the cross-correlation of subsample images to measure and detect blocking artifacts. Eight sub-images are chosen such that every sub-image contains one specific pixel from each of the 88 blocks. The summation of the phase correlations between some sets of sub-images, which measures the intra-block similarity, is divided by the summation of the phase correlations between some other sets of sub-images, which measures the inter-block similarity, to yield a measure of blockiness.

- **M3**, developed by Wang et al. [141], (hereafter also named WBE): It is formulated in the frequency domain and models the blocky image as a non-blocky image interfered with a pure blocky signal. The goal of the blocking effect measurement algorithm is then to detect and estimate the power of the blocky signal. For simplicity, we assume the size of the test image is M x M and the block size is B x B, where M is a multiple of B. An ideal 1-D blocky signal b is defined as

$$b((i+1)B) = b(iB) + V(i)\Delta \quad i = 0, 1, 2, ... \tag{3.3}$$

where $V(i)$ is a random variable that takes on the value of 1 or -1 and Δ is the step size. The blockiness measure should be independent of $V(i)$. To remove the influence of $V(i)$, we first take the absolute difference along the signal:

$$d(i) = |b(i) - b(i-1)| \quad i = 1, 2, ... \tag{3.4}$$

The blockiness measure is therefore defined as the power of the sequence $d(i); i = 0, 1, 2, ...$:

$$M_v = \frac{\Delta^2}{B} \tag{3.5}$$

Thinking the blocky image as a non-blocky image interfered with an ideal blocky signal, the blocking measurement problem is then to detect the blocky signal and estimate its power. Peaks in the spectra due to block structures are identified by their locations. The power spectra of the underlying non-blocky images are approximated by median filtering the aforesaid average power spectra. The overall blockiness measure is then computed as the difference between these power spectra at the peak locations. Luminance and texture masking effects are also integrated within the metric.

- **M4**, developed by Wang et al. [142], (hereafter also named WSB): The method works in the frequency domain and is based on gradient features. It considers blurring and blocking as the most significant artifacts generated during the JPEG compression process. After extracting the features that can be used to reflect the relative magnitudes of these artifacts, these features are combined to constitute a quality prediction model.

- **M5**, developed by Pan et al. [94]: It is based on gradient features and it examines the blocks individually, measuring the severity of blocking artifacts locally. The local metric is averaged over all possible blocks to yield a unique score. It takes into account the blocking artifacts for high bit rate images and the flatness for the very low bit rate images. The authors demonstrate that their method does not require the exact location of the block boundary and is thus invariant to displacements, rotations and scalings of the images.

- **M6**, developed by Muijs and Kirenko [83]: their method allows the block-edge locations and their visibility to be determined. The normalized pixel gradient is calculated and added horizontally or vertically. The presence of blocking artifacts results in pronounced peaks in the summation. The block size and block offset can be extracted by analyzing those peak locations. The blockiness measure is the ratio between the average value at the block-edge locations and the average value at the non-edge locations. Liu and Heynderickx [69] extended the measurement method in [83] by integrating the block-edge grid detecting method from [103] and some masking effects of the HVS.
- **M7**, developed by Chen and Bloom [21]: the method measures the blockiness without any a priori knowledge of the block origin and block size. For a given image, the absolute difference between horizontally adjacent pixels is computed, normalized, and averaged along each column. A one-dimensional discrete Fourier transform is thereafter employed and a vertical blockiness measure is derived. A horizontal blockiness measure is computed similarly. Finally, a blockiness measure for the given image is formulated by pooling those two directional blockiness measures.

3.2 NR metrics for noise

In general, the techniques for estimating the standard deviation of Additive White Gaussian Noise (AWGN) from a single image assume as starting point the following model

$$z(x) = y(x) + \eta(x), \quad \eta(x) \sim N(0, \sigma^2) \tag{3.6}$$

where y is the unknown true image value, η represents the AWGN that corrupts the observed image y and x is a vector representing pixel coordinates.

The algorithms that estimate the standard deviation of a stochastic process η mainly follow a filtering approach. The filtering approach exploits the separation of noise from true image, which is generally obtained subtracting from the (noisy) observation z a smoothed observation obtained by filtering z. This can be done using both linear and non linear filtering such as averaging filters or block-wise median [90]. Other algorithms following this approach have been also introduced: Rank et al. [99] for example propose an algorithm based on Differentiating Filters. Immerkaer [51] introduced a Laplacian mask filtering on the noisy image that allows fast noise variance estimation. The same Laplacian filtering followed by an edge detector has been suggested by Corner et al. [26].

Many other methods exist in the literature but they address multiple distortions like for example noise and blur or noise and JPEG distortions (see for example Table 2.3 of Chapter 2)

Since the Immerkaer [51] algorithm performs well for a large range of noise variance values and it is very simple and fast (requires only the use of a 3x3 mask followed by a summation over the image or a local neighborhood), in what follows we choose it as NR specific method to assess IQ in noisy images.

3.3 NR general purpose metrics

The general-purpose NR IQA algorithms do not attempt to detect specific types of distortions. Methods of this type typically reformulate the IQA problem into a classification and regression problem in which the regressors/classifiers are trained using specific features. The relevant features are either discovered via machine learning or specified by using natural-scene statistics [127, 122, 66, 154].

Another approach to NR IQA is to use natural scene statistics. The main idea in this approach is that natural images demonstrate certain statistical regularities that can be affected in the presence of distortion. Thus, quality can be estimated by extracting features which indicate the extent to which these statistics deviate in the distorted image.

Methods of this type usually contain two stages: (1) distortion identification and (2) distortion-specific quality assessment. Both stages require training: the classifier used to measure the probability that each distortion type exists in the distorted image requires training, and the regression model for each distortion type used to map the measured features to an associated quality score must also be trained. Moorthy and Bovik [81] presented the BIQI algorithm, which estimates quality based on statistical features. Multidimensional feature vectors (3 scales 3 orientations 2 parameters) are used to characterize the distortion and estimate quality via the afore mentioned two stage classification/regression framework. Moorthy and Bovik [80] presented also the DIIVINE algorithm, which improves upon BIQI by using a steerable pyramid transform with two scales and six orientations. The features extracted in DIIVINE are based on statistical properties of the subband coefficients. A total of 88 features are extracted and used to estimate quality via the same two stage classification/regression framework. Saad et al. [104] presented the BLIINDS algorithm which estimate quality based on Discrete Cosine Transform statistics. BLIINDS-I operates on 1717 image patches and extracts DCT-based contrast and DCT-based structural features. DCT-based contrast is defined as the average of the ratio of the non-DC DCT coefficient magnitudes in the local patch normalized by the DC coefficient of that patch. The DCT-based structure is defined based on the kurtosis and anisotropy of each DCT patch.

Mittal et al.[4] presented the BRISQUE algorithm, a fast NR IQA algorithm which employs statistics measured in the spatial domain. BRISQUE operates on two image scales; for each scale, 18 statistical features are extracted. The 36 features are used to perform distortion identification and quality assessment via the aforementioned two-stage classification/ regression framework. It uses scene statistics of locally normalized luminance coefficients to quantify possible losses of naturalness in the image due to the presence of distortions, leading to a holistic measure of quality.

Let us remark that the above cited algorithms (DIIVINE, BLIINDS and BRISQUE) have been trained on a database of human rated distorted images and associated subjective opinion scores. Therefore, these models are called opinion-aware (OA). Given the impracticality of obtaining collections of distorted images with co-registered human scores, models that do not require training on databases of human judgments of distorted images, and hence are opinion unaware (OU), are of great interest. Among these OU models, we can cite the the Natural Image Quality Evaluator (NIQE) developed by Mittal et al. [78]. This method is based on constructing a collection of quality aware features and fitting them to a Multivariate Gaussian (MVG) model. The quality aware features are derived from a simple but highly regular Natural Scene Statistic (NSS) model.The quality of a given test image is then expressed as the distance between the MVG fit of the NSS features extracted from the test image, and a MVG model of the quality aware features extracted from the corpus of natural images.

3.4 Correlating objective and subjective data

In order to evaluate the goodness of a chosen metric in predicting the subjective quality scores, it is customary to apply a regression function to correlate the subjective and objective data. As already mentioned in Section 2.5, metrics that highly correlate with human ratings typically yield high PCC and SROCC correlation coefficients.

In Figures 3.1 and 3.2 we plot some of the above metrics' results versus the Differential Mean Opinion Scores (DMOS) for the JPEG LIVE database. The logistic regression curves are shown. In Table 3.1 the corresponding PCC and SROCC coefficients are reported.

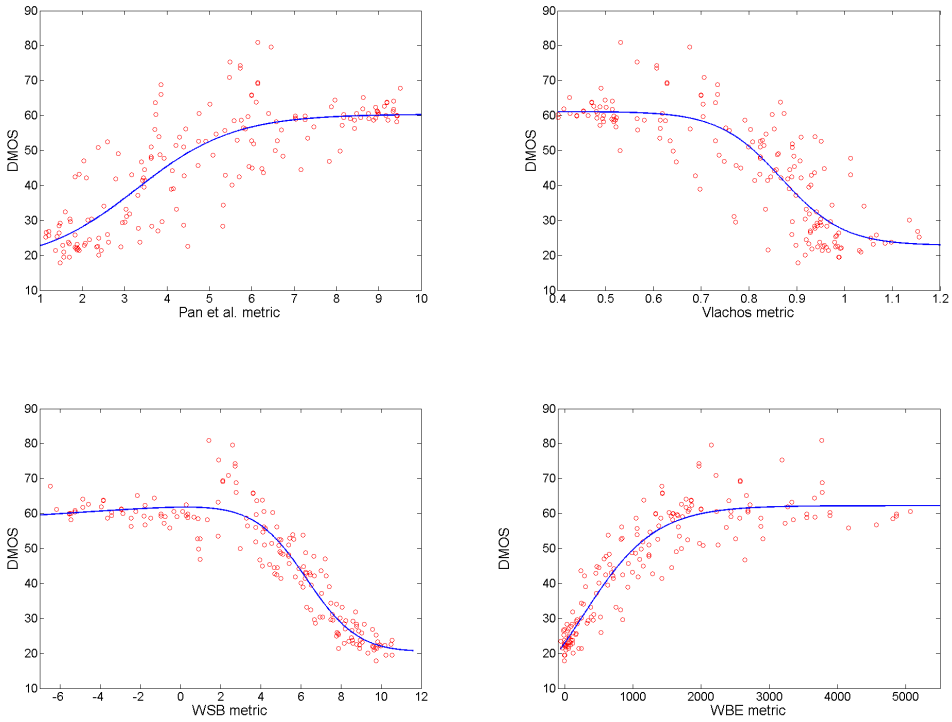

Figure 3.1: Logistic regression for different metrics for JPEG distortion. First row: Pan [94] and Vlachos et al. [135]; second row: WSB [142] and WBE [141]. Regression is performed on LIVE database for JPEG distortions.

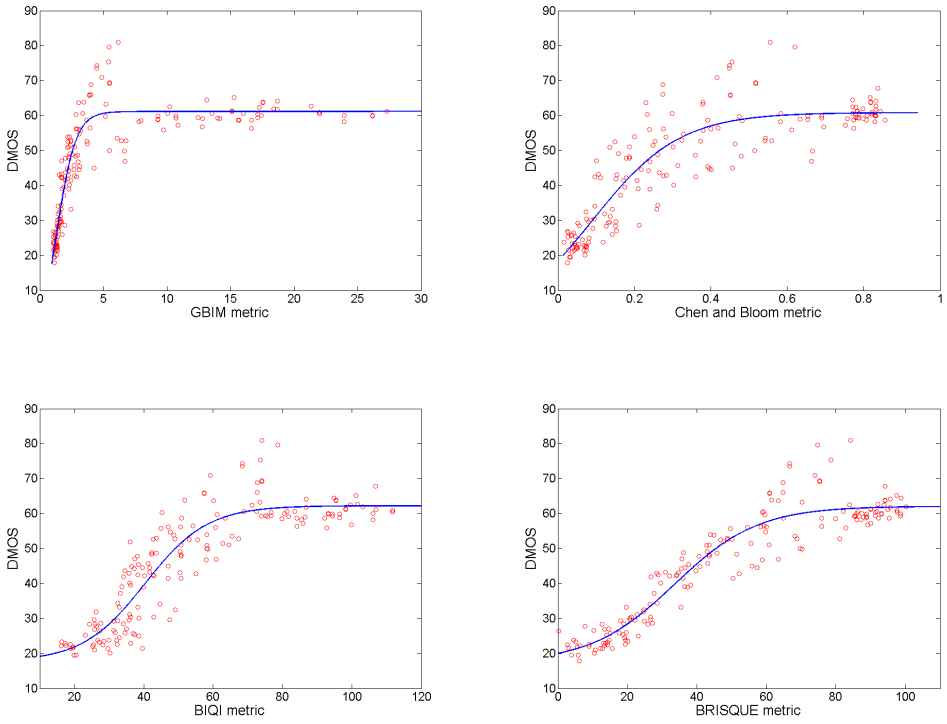

Figure 3.2: Logistic regression for different metrics for JPEG distortion. First row: GBIM [152] and Chen and Bloom [21]; second row: BIQI [81] and BRISQUE [4]. Regression is performed on LIVE database for JPEG distortions.

Table 3.1: PCC and SROCC for NR metrics on JPEG LIVE database.

Correlation	**Pan**	**Vlachos**	**WSB**	**WBE**	**GBIM**	**Chen**	**BIQI**	**BRISQUE**
PCC	0.8173	0.8407	0.9373	0.9216	0.9319	0.8924	0.9065	0.9478
SROCC	0.8006	0.8304	0.8789	0.8786	0.8931	0.8614	0.8886	0.9161

We observe from Table 3.1 and Figures 3.1-3.2 that, even if PCC and SROCC coefficients are high, within each single metric we can find some images with significantly different subjective scores but corresponding to similar values of the metric. Let us consider for example the metric by Chen and Bloom [21] and the two images shown in Figures 3.3 from the JPEG LIVE database. We note that the DMOS value equal to zero indicates the reference (best quality) while DMOS = 100 indicates the worst quality. The objective metric by [21] varies from zero (best quality) to one (worst quality). The image on the top of Figure 3.3 has been subjectively evaluated with a a DMOS of 45 and the metric's prediction is equal to 0.46. The figure on the middle corresponds to a DMOS of 75 and the metric's prediction value of 0.45. Finally, on the bottom we plot again the regression curve where both images are highlighted.

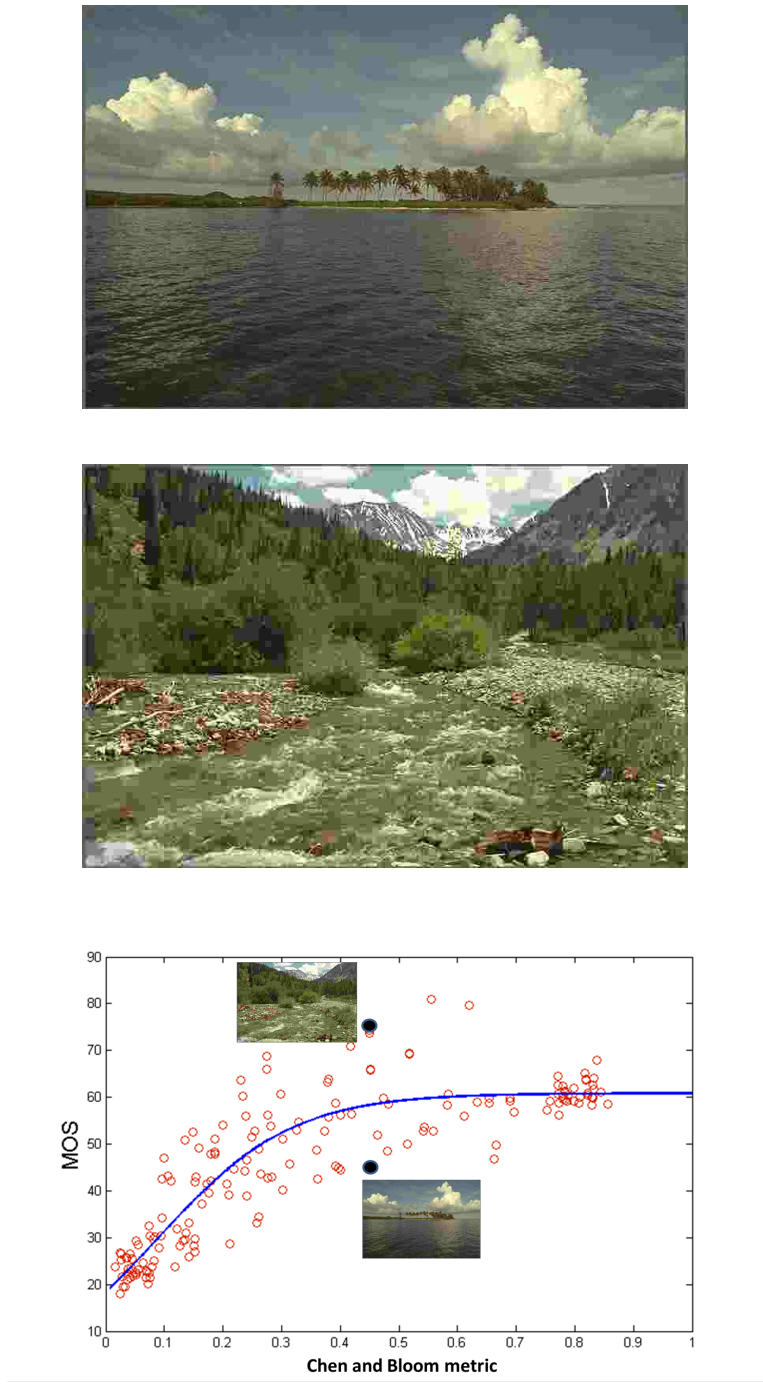

Figure 3.3: Two images from JPEG LIVE database. The image on the top has been subjectively evaluated with a a DMOS of 45 and the metric value by [21] is equal to 0.46. The figure on the middle corresponds to a DMOS of 75 and metric value of 0.45. On the bottom the regression curve is again plotted where both figures are highlighted.

This behavior can be also observed in the other NR JPEG-blockiness metrics as well, and it is typically found in the intermediate region of the distortion range. That is, this data dispersion decreases as we focus on images slightly distorted or highly distorted. To illustrate this, we consider images as the the combining of two signals: content and distortion. As the distortion increases, the visibility of the content decreases. We can thus locate images within a plot where the amount of content and distortion are taken into account as in Figure 3.4. High quality images (like for example those acquired by professional camera) are placed on the left portion of this graph. Their content is dominant with respect to the distortions. In the right portion of the plot, we locate the images where the distortions are so significant that the content is recognized with difficulty and when applying a metric, we reasonably measure the distortion itself. In the intermediate range both content and distortion are significantly present and consequently, not easily decorrelated to be measured. We observe that, in general, NR metrics are not able to measure with the same performance the distortions within their possible range and with respect to different image contents. Moreover, the crosstalk between content and distortion signals influences both the subjective and objective quality assessment. We address these two issues in Chapters 5 and 6 respectively.

Let us also note that, within this point of view, Larson and Chandler [64] (see also references therein), claim that our visual system uses different strategies to evaluate image quality depending on the signal-distortion ratio. In the high quality regime, the visual system attempts to look for distortions in the presence of the image, whereas in the low quality regime, the visual system attempts to look for image content in the presence of the distortions. Based on this hypothesis, the authors propose a FR method which attempts to explicitly model these two separate strategies.

Following this philosophy, in Chapter 5 we will approach the IQA as a classification problem where our target is the identification of the following three classes: images where the distortion is not perceived (content is dominant with respect to the distortion), images where the artifacts are easily observed (distortion more or as significant as the content signal), and images where an observer has doubt about the presence or not of the artifact (clear interference between content and distortion signal).

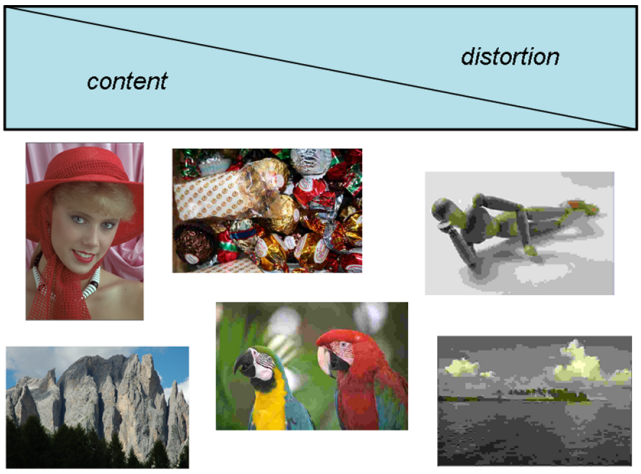

Figure 3.4: On the left portion of the graph we position high quality images, for which the content is dominant with respect to the distortions. On the right portion of the plot we position low quality images for which the distortion is dominant with respect to the image signal. In between we find images for which content and distortion are strongly correlated.

Correlation for noise data

In Figure 3.5 we correlate NR metrics and DMOS for the noise distorted images of LIVE. Five metrics are considered: Immerkaer metric [51], specific for noise distortion, and four general purpose methods: BIQI, BRISQUE, BLIINDS and NIQE. In Table 3.2 we report PCC and SROCC coefficients for each of these metrics. In this case we observe that all the metrics describe the subjective scores with high performance.

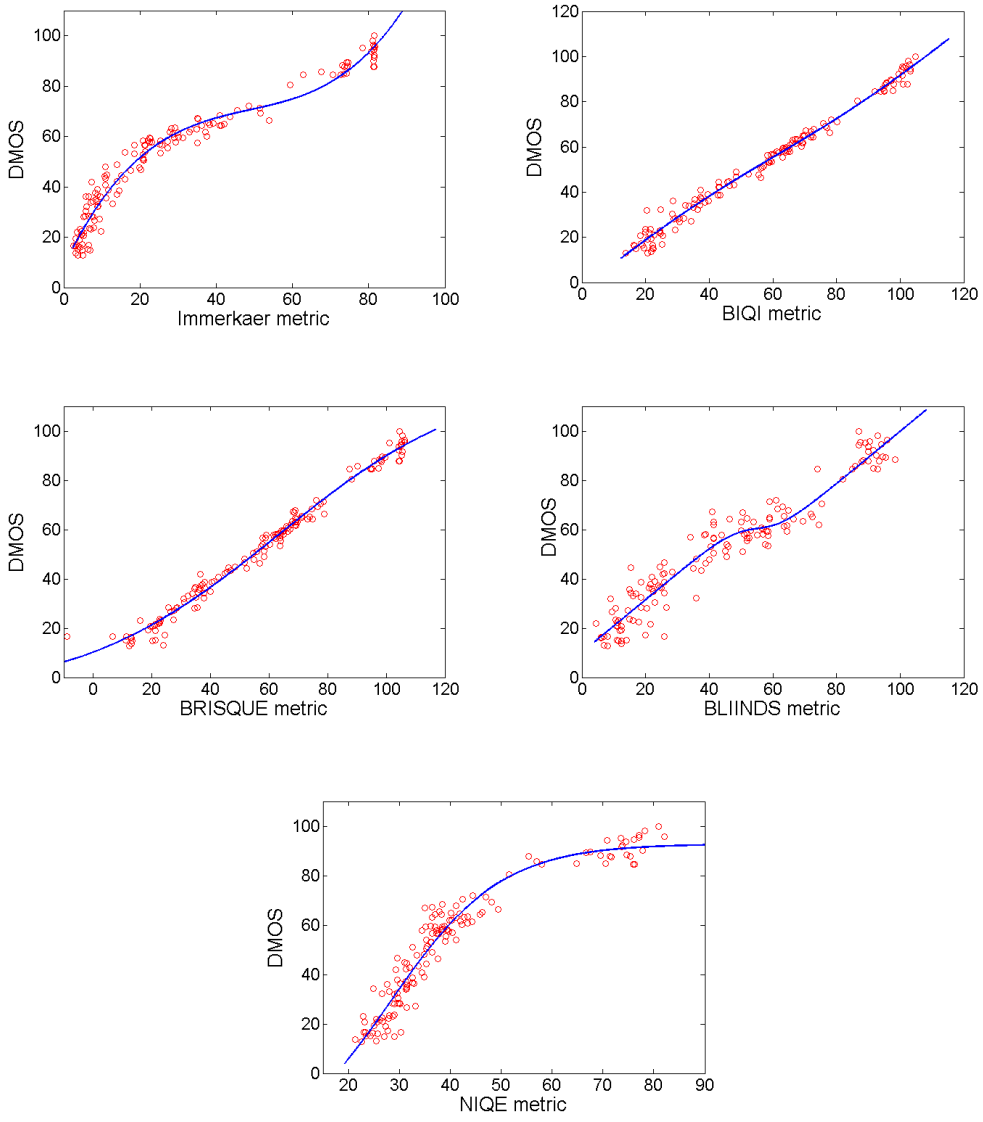

Figure 3.5: Logistic regression for different NR metrics for noise on LIVE data. First row: Immerkaer [51] and BIQI; second row: BRISQUE and BLIINDS; third row: NIQE.

Table 3.2: PCC and SROCC for NR metrics on noise LIVE database.

Correlation	**Immerkaer**	**BIQI**	**BRISQUE**	**BLIINDS**	**NIQE**
PCC	0.9809	0.9930	0.9926	0.9652	0.9656
SROCC	0.9793	0.9903	0.9911	0.9496	0.9544

Chapter 4

The IVL database

4.1 Generation of the IVL database

In this chapter we introduce the Imaging and Vision Lab (IVL) database, generated with the aim of assessing the quality of images corrupted by JPEG or Gaussian noise.

The IVL database starts from 20 original images of 886x591 pixels (15x10 cm at 150 dpi, typical printing parameters for natural photos), chosen to sample different contents both in terms of low level features (frequencies, colors) and higher ones (face, buildings, close-up, outdoor, landscape). The corresponding thumbnails are shown in Figure 4.1.

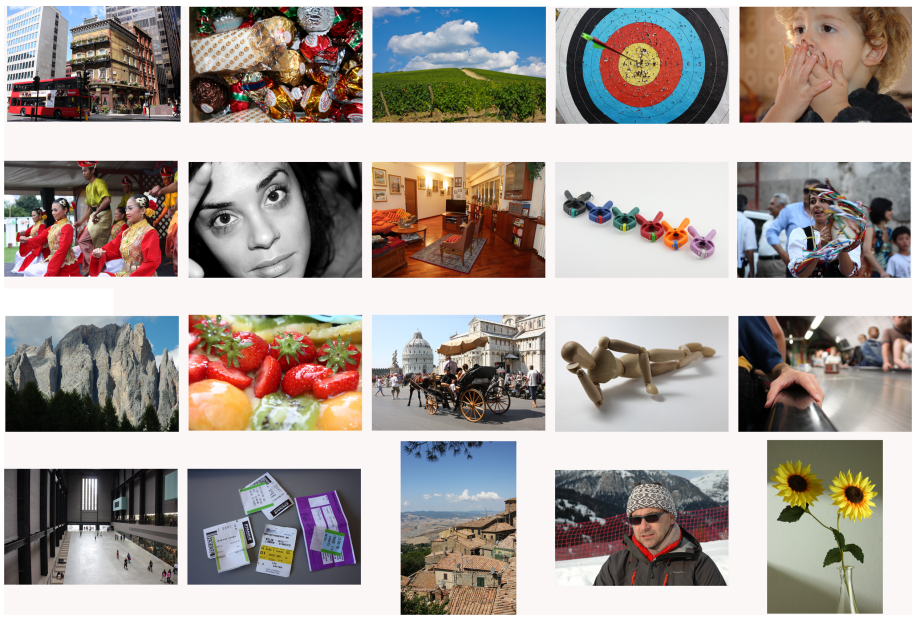

Figure 4.1: The 20 original images of the IVL database.

With respect to the JPEG distortion, the compressed images were generated using the Matlab imwrite function. As the Q-factor depends on each different JPEG compression algorithm, we have adopted the bit per pixel (bpp) Ratio (*bppR*) with respect to a reference, finding iteratively the Q-factors that better match the corresponding *bppR* values.

As reference we have adopted the $Q = 100$ compressed image, where the compression is mainly due to the sub sampling of the chroma channels and to lossless algorithms. For each of the 20 original images, we have created 9 compressed versions with the following *bppR*: $1(Q = 100), 0.707, 0.5, 0.25, 0.177, 0.125, 0.105, 0.088, 0.0625$. We have chosen these values empirically, sampling an exponential function, to include 1, 1/2, 1/4 and 1/8 and to represent perceptually significant variations in the JPEG artifacts. The exponential function and the chosen *bppR* values are reported in Figure 4.2. The final JPEG database is composed by 20x9 = 180 distorted images.

With respect to the noise distortion, each of the 20 original images of the IVL database were corrupted with Gaussian noise on the luminance channel. For each image we created 10 corrupted versions with $\sigma = 1, 2, 3, 4, 5, 6, 8, 10, 12, 14$ Gray Level of Standard Deviation (GLSD). The final noise database is composed by 20x10 = 200 distorted images.

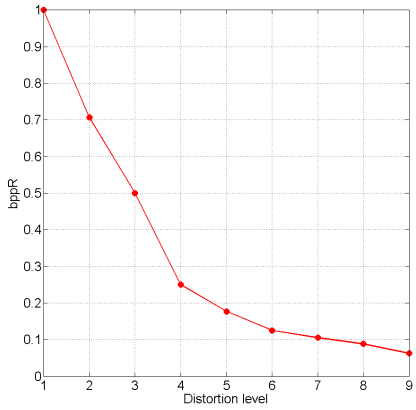

Figure 4.2: The exponential function from where the 9 *bppR* that sample the JPEG distortion were extracted; and used to generate distorted images of the IVL database.

4.2 Test conditions and experimental methods

To perform the psycho-visual tests, the images were shown on a web-based interface. A Javascript slider assigning a quality score was used. The workstations adopted were placed in an office environment with normal indoor illumination levels. The ambient light levels were maintained constant between the different sessions. We adopted mid-range LCD monitors properly calibrated with a colorimeter (D65, gamma 2.2) [7, 116].

For the quality analysis of the images of each of the two databases (JPEG and NOISE), we have adopted a Single Stimulus method (SS) [ITU02]. Usually, when it is important to check the fidelity with respect to the source signal, DS method should be used. Even if the reference images are available in the present study, in our experiments we have decided to adopt the SS method to better represent the reality where users of digital photographs do not generally dispose of the reference image (NR image quality assessment).

In all our experiments, distorted images are shown in a random order, different for each subject. The subjects report their quality judgments by dragging a slider onto a quality scale. The position of the slider is automatically reset after each evaluation. Observers were asked to provide their perception of quality on a continuous linear scale with three reference labels. The continuous scale has been mapped into three regions associated to three classes: high quality, middle quality, and low quality; as shown in Figure 4.3.

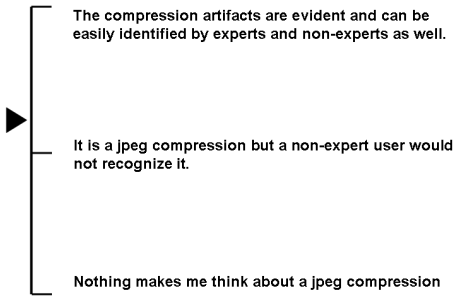

Figure 4.3: Quality scale slider for the JPEG test. A similar one is used for the NOISE test.

The panel of subjects involved in this study was recruited from our laboratory and from Océ [3]. The subjects involved were experienced with image quality assessment and image impairments. The total number of subjects was 31.

Each image including the original is evaluated according to the following steps:

- At the beginning of each cycle of the test, a synthetic image, showing geometric features, is presented. This image permits to reset our visual system with respect to previously analyzed images.
- The first image is shown and remains visible on the screen for a time in seconds indicated as MAXt.
- The image disappears and the quality scale appears and remains till the subject makes his judgment.

The values of MAXt have been estimated during what we call the tuning tests, described in the following. These four steps procedure of the subjective test is sketched in Figure 4.4.

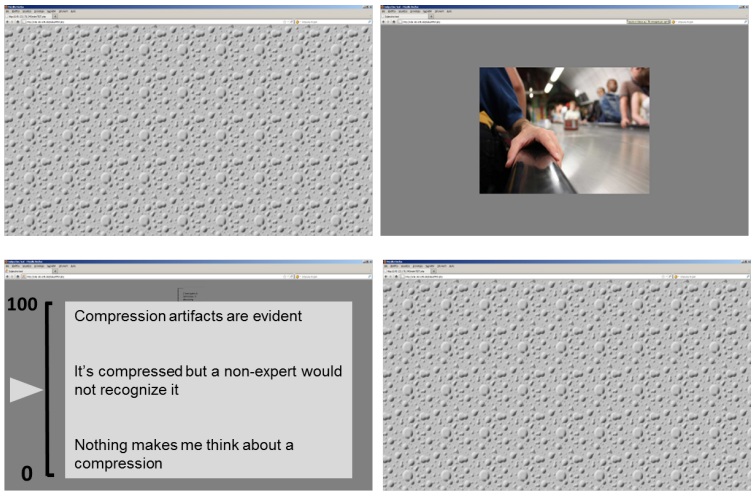

Figure 4.4: Four steps procedure of the Single Stimulus method employed during the experimental sessions.

4.3 Experimental sessions

In order to collect the subjective data, we have performed different experimental sessions: tuning, preliminary and test sessions, as summarized in Table 4.1.

Table 4.1: Experimental sessions

Session	Description	Sessions No	Images No
Tuning	common to JPEG and NOISE	1	4x4
Preliminary JPEG	Implicit training using SS method	1	4x4
Preliminary NOISE	Implicit training using SS method	1	4x4
Test JPEG	Efffective test using SS method	5	4x9
Test NOISE	Efffective test using SS method	5	4x10

4.3.1 Tuning sessions

Before starting the preliminary and test sessions, an initial analysis of the test structure and organization has been performed to better tune the successive experiments. In particular, with this tuning session we have studied the test efficacy and the best way to perform the experiments. In particular we have obtained the best visualization time for each image and the maximum duration of the whole experiment for each participant. With respect

to 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. This tuning session is common for both on the JPEG and NOISE databases.

4.3.2 Preliminary sessions

During a preliminary test, each subject is implicitly trained about the nature and range of the distortion to be evaluated. These preliminary sessions aim to avoid that this training occurs during the effective test, conditioning the experimental results. In this way, the MOS values collected result uniformly distributed across the entire range. We have preliminary sessions for all the subjects involved and for each of the experiments. In the Tables 4.2 and 4.3 we summarize the JPEG and NOISE preliminary sessions carried out.

Table 4.2: JPEG preliminary session

subjects	6
Maxt	10 seconds
Visual adaptation image	2 seconds
Images	4 references x 4 distorted
compression factors	$bppR = 1, 0.125, 0.088, 0.0625$

Table 4.3: NOISE preliminary session

subjects	6
Maxt	10 seconds
Visual adaptation image	2 seconds
Images	4 references x 4 distorted
distortion	Gray level of standard deviation = 1, 3, 10, 14

4.3.3 Test sessions

The details of the test sessions are summarized in Tables 4.4 and 4.5

Table 4.4: JPEG test sessions

Number of sessions	5 days (Monday to Friday)
subjects	31
Maxt	10 seconds
Effective test duration	10-15 minutes
Images	36 images per session

Table 4.5: NOISE test sessions

Number of sessions	5 days (Monday to Friday)
subjects	31
Maxt	10 seconds
Effective test duration	10-15 minutes
Images	40 images per session

4.4 Data analysis

For each of the 180 JPEG-degraded images and for each of the 200 noisy images, we have collected the continuous values assigned by all the participants and the corresponding MOS have been calculated. Keeping in mind one of the goals of the present research, i.e. to classify images according to their quality, we have also converted each of the single continuous values into three classes dividing uniformly the entire scale. The final class assigned to a given image is the class with highest frequency among all the viewers. Using the statistical mode we discard the influence of outliers.

In Figure 4.5 the final classes of the 180 JPEG degraded images are shown with respect to the level of distortion in terms of $bppR$. Each column sums up to 20 (the number of reference images) and reports the proportion of the assigned classes. Images with the highest level of compression ($bppR = 62$ permil) are all in class 1 (first column), independently of their content. On the other hand, class 3 spans a larger number of distortion levels (from $bppR = 1$ to $bppR = 177$ permil), corresponding to the last 5 columns. This fact is not unexpected since JPEG compression aims to preserve as much as possible the perceptual quality, even with strong compression levels.

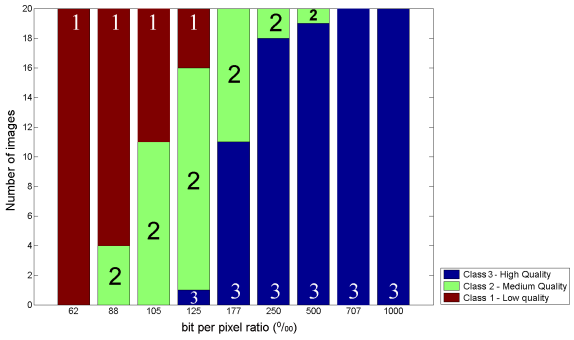

Figure 4.5: Bar diagram of the psycho-visual data for JPEG distortion.

In Figure 4.6 the subjective classes of the 200 noisy images are shown with respect to the noise level in units of GLSD. As in the JPEG case, each column sums up to 20 (the number of reference images) and reports the proportion of the assigned classes. Images with the minimum amount of noise are all in class 3 (independently of the content) while images with the highest level of noise are all in class 1 (independently of the content).

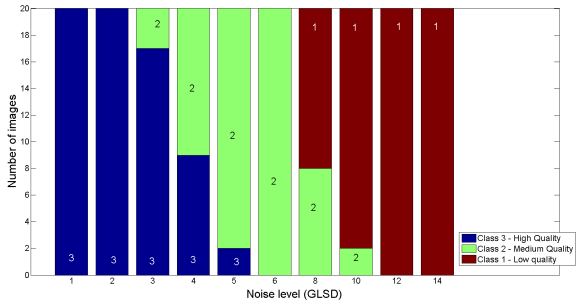

Figure 4.6: Bar diagram of the psycho-visual data for noise distortion.

4.4.1 Correlation of the IVL data for JPEG-blockiness

In this section the regression between different NR metrics and the MOS corresponding to IVL data is performed. In Figures 4.7-4.8 the subjective data is correlated with six NR metrics specifically developed for JPEG-blockiness [94, 142, 141, 152, 83, 21] and two general purpose ones [4, 78], using the logistic regression function. In Table 4.6 the correlation coefficients PCC and SROCC are reported for these metrics.

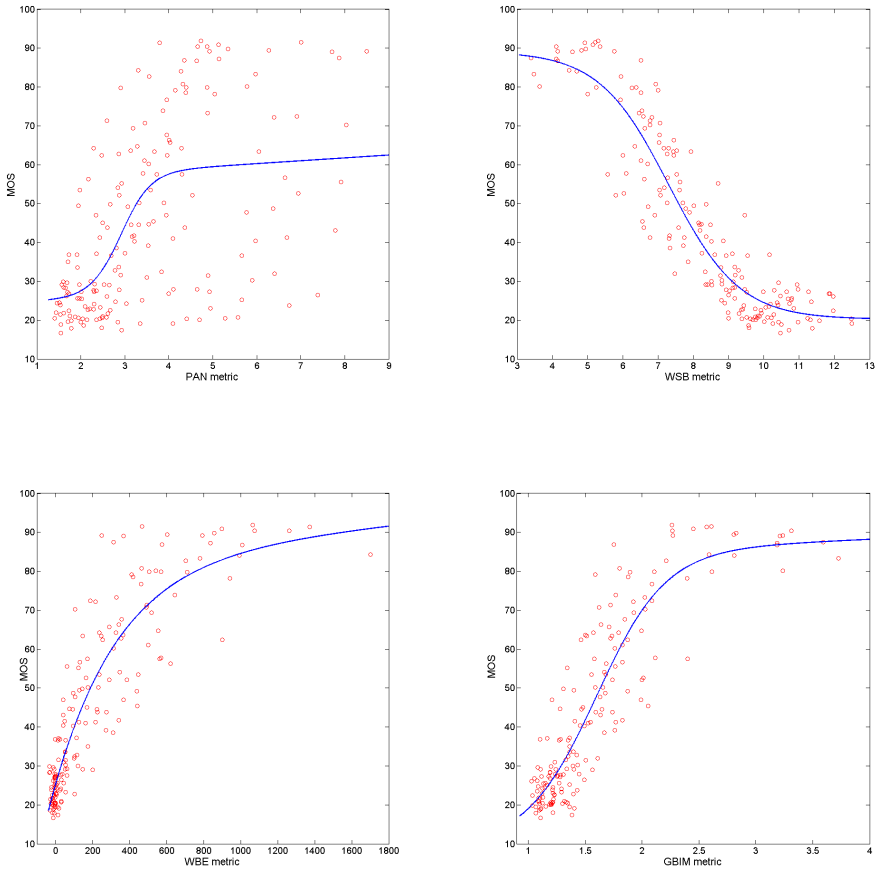

Figure 4.7: Logistic regression for different metrics for JPEG distortion on IVL data. First row: Pan [94] and WSB [142]; second row: WBE [141] and GBIM [152].

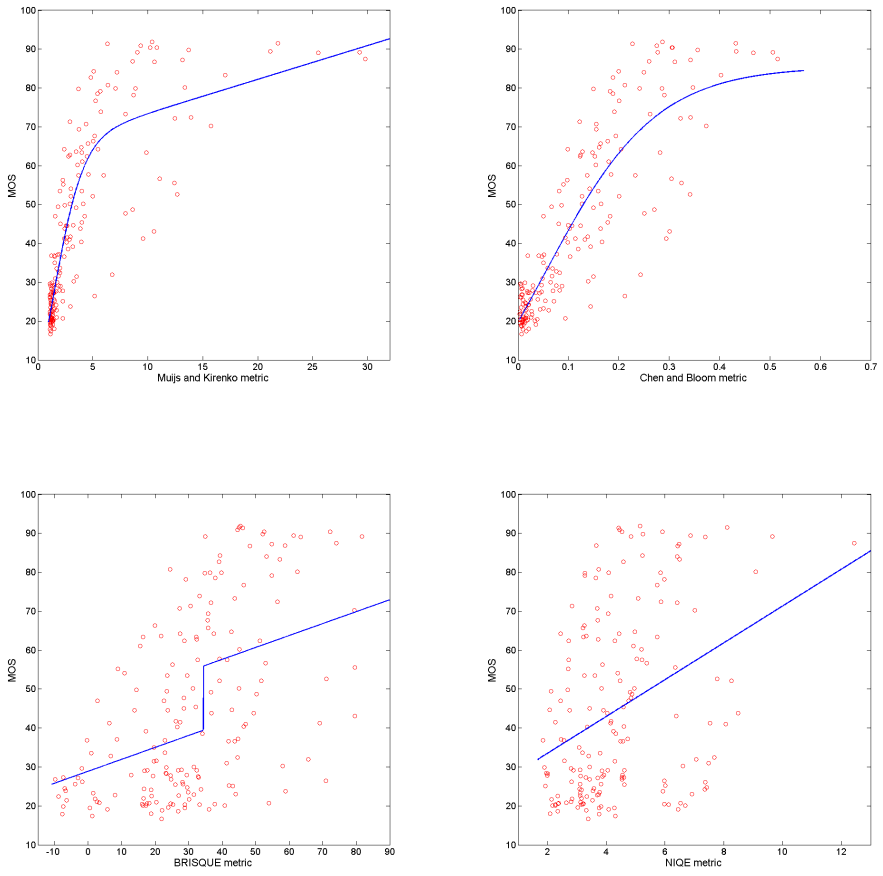

Figure 4.8: Logistic regression for different metrics for JPEG distortion on IVL data. First row: Muijs and Kirenko [83] and Chen and Bloom [21]; second row: BRISQUE [4] and NIQE [78].

Table 4.6: Pearson and Spearman correlation coefficients for NR JPEG-blockiness and general purpose metrics on IVL database.

Correlation	**Pan**	**WSB**	**WBE**	**GBIM**	**Muijs**	**Chen**	**BRISQUE**	**NIQE**
PCC	0.6012	0.9302	0.9059	0.9028	0.8789	0.8685	0.5747	0.3593
SROCC	0.5686	0.9042	0.8922	0.8662	0.8714	0.8689	0.5387	0.3534

Comparing the correlation indexes of Table 4.6 with respect to those of Table 3.1 (that correspond to the regression with LIVE data), we observe that some of the metrics show lower performances. In particular, both general purpose metrics are the ones with smallest correlation indexes. We recall that the metric BRISQUE has been defined and trained on LIVE data. Also the performance of the JPEG-blockiness specific metric by Pan [94]

is smaller. This decrease of performance might be partially explained by the difference of quality range in the databases. Low quality anchors in the LIVE database are indeed strongly distorted pictures with extremely low quality. The corresponding low anchors in the IVL database have a much better quality. Similar conclusions have been already pointed out by Tourancheau et al. [131] who studied the impact of subjective dataset on the performance of FR IQ metrics. The authors compared the behaviour of FR metrics (PSNR, SSIM, VIF) on three image databases (LIVE, IVC, MICT) and they demonstrated that the performances of the quality metrics can strongly fluctuate depending on the database used for testing. We will further analyze this issue in Chapter 5.

Considering each of the NR metrics separately, and focusing on a single image (i.e. single content), in general a monotone behavior is observed as the blockiness increases. For example, in Figure 4.9 the absolute value of the **M3** metric [141] is plotted as function of the distortion factor for five example images of the IVL database. Focusing on only one image (corresponds to one curve), we observe that the metric decreases as the blockiness increases as desired. However, if different image contents are considered a non monotone profile is obtained. This behavior is due to the fact that different contents influence differently the measure of the same level of distortion. We will further analyze this issue in Chapter 6.

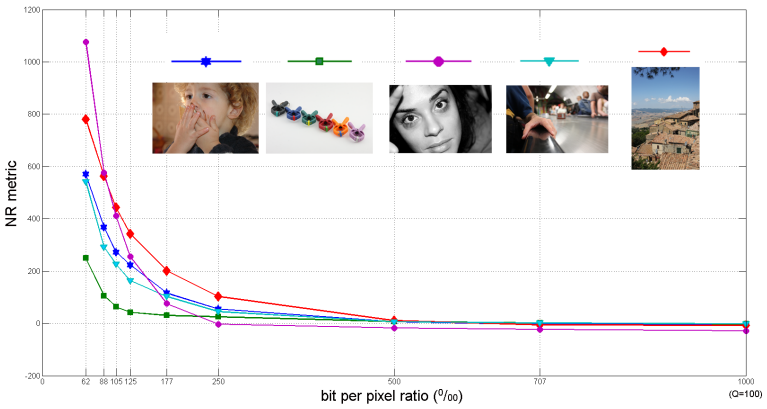

Figure 4.9: The **M3** metric [141] applied to 5 different images of the IVL database with 9 levels of JPEG compression.

4.4.2 Correlation of the IVL data for noise

In Figure 4.10 the metric by [51] is plotted as function of the noise distortion level in units of GLSD. Twenty different curves are shown where each of them correspond to each of the original images and their ten distorted versions. In Figure 4.11 the subjective scores corresponding to the noise IVL database are shown as function of the objective metric values and also the logistic regression curve is reported. In this case, the Pearson correlation coefficient results equal 0.9695.

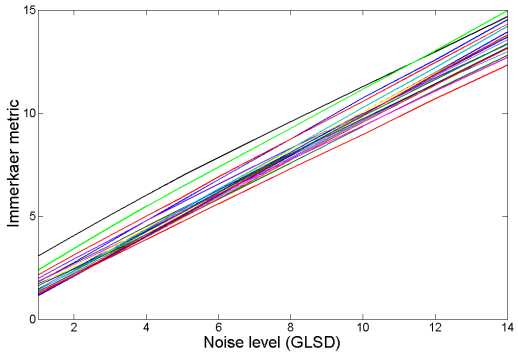

Figure 4.10: Noise metric by [51] applied to the IVL database as a function of the distortion level. Twenty curves are observed, each of them corresponding to each of the twenty original images.

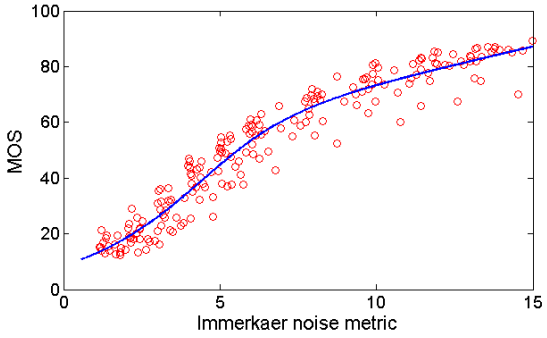

Figure 4.11: Logistic regression for the Immerkaer metric and the noise IVL database.

Similarly, the behavior of the general purpose metrics as function of the noise distortion level is shown in Figure 4.12 for BRISQUE (left) and NIQE (right) metrics. In this case we obtain PCC = 0.9263 and PCC = 0.7396 respectively. The corresponding logistic regression curves are depicted in Figure 4.13.

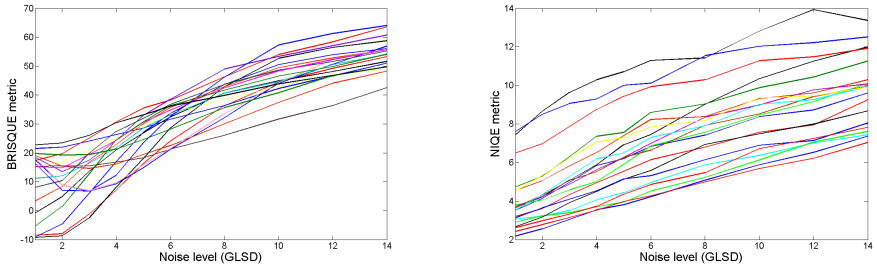

Figure 4.12: General purpose metrics applied to the noise IVL database as function of the noise distortion level. Left: BRISQUE metric, right: NIQE metric. Twenty curves are observed, each of them corresponding to each of the twenty original images.

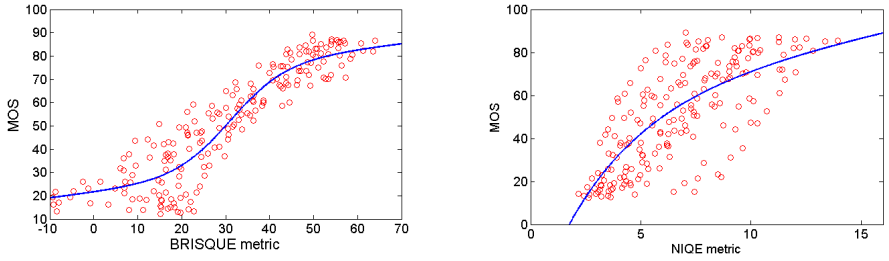

Figure 4.13: Logistic regression for BRISQUE (left) and NIQE (right) metrics and the noise IVL database.

For the case of noise on IVL data, the simple metric by [51] is the one with highest correlation.

Chapter 5

Image quality assessment as a classification problem

In Chapters 2 and 3 many NR methods were summarized: those targeted to estimate the presence of a specific single image defect (bluriness, JPEG-blockiness, grainness, noise, colorfulness) and also the blind ones, where no information about the distortion affecting the images is known. Moreover two sub-categories of general purpose methods have been identified: the Opinion Aware (OA) models, that are those trained on a database of human rated distorted images and associated subjective scores, [81, 4, 104], and Opinion Unaware (OU) ones that do not require such training [78]. In Chapter 3 we mainly focussed on JPEG-blockiness metrics and we have observed that IQA metrics are not in general able to measure the distortions with the same performance within their possible full range and with respect to different image contents. Also psycho-visual experiments have shown that the perception of distortions is influenced by the amount of details in the images content [7].

Given the variety of available objective metrics and databases, some questions arise: how do the metrics behave across different databases? Given a distortion type and several NR metrics, which of the metrics best measures the distortion? Does a combining of the NR metrics improve the single methods? In this chapter we try to elucidate in particular the last issue concerning the combining for the case of NR metrics and JPEG artifacts. To do so, the way we propose here is to approach the NR-IQA field by focusing on a classification problem. That is, we aim to investigate a methodology that let us evaluate how JPEG (or noise) corrupted images can be classified within different groups or classes, according to their quality.

To solve this classification problem, we consider two different machine learning methods: Classification and Regression Tree (CART) [13] and Support Vector Machine (SVM) [134].

In the last years, machine learning methodologies have been applied within the field of IQA [85, 84, 20, 126, 71, 120]. Both Charrier et al. [20] and Liu et al. [71] focus on FR methods. The approach by Charrier et al. uses multi-Support Vector Machine classification, where the quality classes are according to the quality scale recommended by the ITU. To evaluate the quality of images, a feature vector containing visual attributes describing images content is constructed. Then, a classification process is performed to provide the final quality class of the considered image. Finally, once a quality class is associated to the considered image, a specific SVM regression is performed to score its quality. Liu et al. [71] propose a methodology for IQA with multi-metric fusion. Using a

machine learning regression approach, the authors fuse multiple FR metrics and demonstrate that it is possible to achieve significantly better performance at the cost of higher complexity. The multi-metric fusion score is set to be the nonlinear combination of multiple metrics with suitable weights obtained by a training process. Narwaria and Lin [84] propose an IQA algorithm based on support vector regression where the input features are the singular vectors out of singular value decomposition. Suresh et al. [120] present a machine learning approach to measure the visual quality of JPEG-coded images, considering various human visual characteristics. The functional relationship between the extracted features and the subjective scores is modeled by Extreme Learning Machine (ELM) algorithm. The authors transform the problem of quality estimation into a classification problem and solve it using ELM.

An overview of the benefits that the use of machine learning can bring to the visual quality assessment problem can be found in Gastaldo and Redi [40]. The authors also illustrated a number of good practices to set-up a machine learning-based quality assessment system and they exemplified and motivated those practices with a case study, the RR quality assessment system proposed by Redi et al. [100].

In the above cited articles, image features like for example contrast, luminance and wavelets coefficients among others, are used to construct the feature space. We differentiate from those approaches since our proposal is to define a feature space with different NR metrics. In the supervised learning phase of the classification we use the subjective scores obtained from a psycho-visual experiment properly designed.

Addressing the IQA as a classification problem as here presented could be useful in many application domains. For example a three classes classifier could be integrated within a printing workflow chain: images classified as very low quality will be rejected and not printed, images of very good quality will be directly printed, while the rest of the images will be forwarded to the the human judgment. Another example could be a web based image retrieval application where it could be helpful to automatically recover only images of a specific class quality, in particular when dealing with huge databases.

5.1 General framework

A schematic overview of the IQ classification approach proposed here, is shown in Figure 5.1.

To solve this classification problem, we have considered both CART [13] and SVM [134] methodologies. For what concerns the number of classes, following the ITU recommendations we first consider five classes corresponding to the five categorical attributes: *Bad*, *Poor*, *Fair*, *Good*, and *Excellent*. As a second possibility we investigate the case of three classes corresponding to low, medium, and high quality images.

The present classification scheme has been applied to the case of JPEG distortion. We have considered eleven NR metrics, seven specific to measure JPEG-blockiness and four general purpose. The subjective scores are obtained from the psychovisual experiments conducted to assess IQA on different available databases and also on the IVL database introduced in Chapter 4.

5.2 Machine learning methods

Briefly, in the CART methodology the classifiers are produced by recursively partitioning the feature space, each split being formed by conditions related to the feature values. In tree

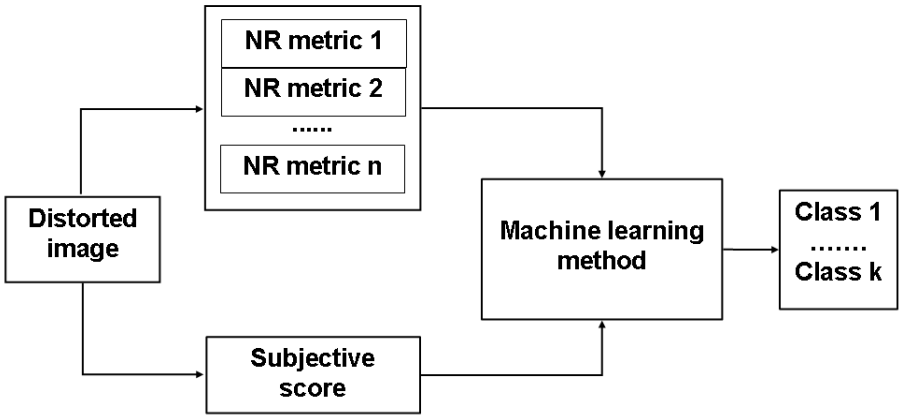

Figure 5.1: Overview of our IQ classification task.

terminology, subsets are called nodes: the feature space is the root node, terminal subsets are terminal nodes, and so on. Once a tree has been built, a class is assigned to each of the terminal nodes, and when a new case is processed by the tree, its predicted class is the class associated with the terminal node into which the case finally moves on the basis of its feature values. The detailed description of the CART methodology can be found in [13]. In our experiments we have used the Matlab implementation. Using CART, it's easy to understand what features are important in making the prediction. Infact, the decision tree generated uses only the features that help to separate the classes, while the others are not considered. In this way CART classifiers provide fairly comprehensible predictors in situations where there are many variables which interact in complicated, nonlinear ways [108, 109, 98]. In this paper we use the tree obtained by a CART not only to solve our classification problem but also as a feature selection method that eliminates redundant information and thus reduces the dimensionality of the feature space.

The SVM methodology comes from the application of statistical learning theory to separating hyperplanes for binary classification problems [134, 27]. The central idea of SVM is to adjust a discriminating function so that it makes optimal use of the separability information of boundary cases. Given a set of cases which belong to one of two classes, training a linear SVM consists in searching for the hyperplane that leaves the largest number of cases of the same class on the same side, while maximizing the distance of both classes from the hyperplane. When the training set is not linearly separable, the optimal separating hyperplane is found by solving a constrained quadratic optimization problem. Although SVMs are mainly designed for the discrimination of two classes, they can also be adapted to multi-class problems. A multi-class SVM classifier can be obtained by training several classifiers and combining their results.

5.3 JPEG-blockiness IQA classification

Eleven different metrics for JPEG distortion, previously described in Chapter 3 are considered: seven specific to measure the JPEG distortions (metrics 1 to 7) and four general purpose (metrics 8 to 11). We name them as follows:

1. **M1**: GBIM metric developed by Wu and Yuen [152]

2. **M2**: developed by Vlachos [135]
3. **M3**: WBE metric developed by Wang et al. [141]
4. **M4**: WSB metric developed by Wang et al. [142]
5. **M5**: developed by Pan et al. [94]
6. **M6**: developed by Muijs and Kirenko [83]
7. **M7**: developed by Chen and Bloom [21]
8. **M8**: BIQI metric developed by Moorthy and Bovik [81]
9. **M9**: BRISQUE metric developed by Mittal et al. [4]
10. **M10**: BLIINDS metric developed by Saad et al. [104]
11. **M11**: NIQE metric developed by Mittal et al. [78]

Comparison of available subjective data for JPEG IQA

Besides the subjective scores obtained from the JPEG IVL data, we have also analyzed several datasets available in the literature that include JPEG distorted images and the corresponding psycho-visual data. We summarize in Table 5.1 these databases only for what concerned the JPEG distortion.

Testing our classifiers on psycho-visual data obtained with different experiments has to be done carefully. Subjective evaluations are influenced by all the aspects of the experimental setup. In particular, for a given defect, they strongly depend on:

- the experimental methodologies (single stimulus, double stimulus, pair wise comparison,...)
- the evaluation scales (continuous scale, adjectival categorical judgment,...)
- the distortion range considered.

Among all the datasets available (see Table 5.1), we find that only LIVE, MICT and IVL can be considered within a classification task as they are single stimulus and provide absolute category ratings. In the LIVE database the 29 original images have been differently JPEG corrupted with about 8 level of distortions for each original image, for a total of 175 distorted images. In the LIVE experiments, observers were asked to provide their perception of quality on a continuous linear scale that was divided into five equal regions, marked with adjectives (*Bad*, *Poor*, *Fair*, *Good*, and *Excellent*). The scale was then converted into 1-100 linearly.

With respect to the MICT database, the 14 original images have been corrupted with 7 levels of distortions, for a total of 98 distorted images. The seven distortions correspond to the following seven levels of Q-factors: $100, 79, 57, 37, 27, 20$, and 15. In the MICT experiment, the subjects were asked to provide their perception of quality on a discrete quality scale corresponding to the 5 categorical adjectives: *Bad*, *Poor*, *Fair*, *Good*, and *Excellent*.

In order to compare the distortion range of these datasets, we consider the *bppR*, that is the bit per pixel (bpp) Ratio between the bpp of a distorted image with respect to the bpp of a lossless compressed version of the original image. In Figure 5.2 the histograms of the *bppR* for LIVE, MICT and IVL are shown. We assigned to the original images

used during the psycho-visual experiments the value of $bppR = 1$ as they are qualitatively equivalent to images compressed with lossless algorithms. Note that LIVE distortions are more concentrated in the range of low $bppR$, with respect to MICT and IVL databases.

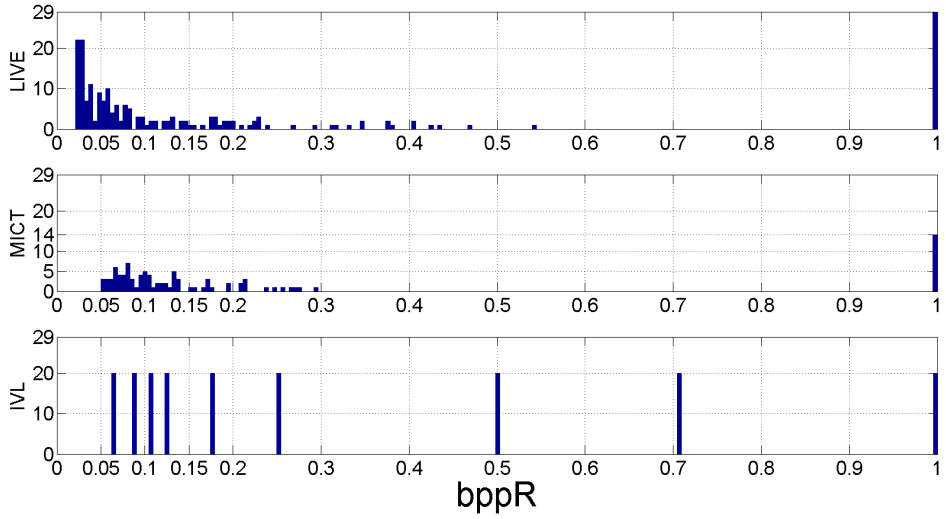

Figure 5.2: Comparison of the $bppR$ histograms of LIVE (top row), MICT (middle row) and IVL (bottom row) databases.

As a qualitative example of the different range of distortions considered, we report in Figure 5.3 one of the most compressed images of LIVE(top), MICT (middle) and IVL (bottom) databases, with $bppR$ equal to 0.0286, 0.05 and 0.062, respectively. The lowest quality images in the LIVE database are much more distorted than the lowest ones in the other two databases considered.

Figure 5.3: The most compressed images of the the LIVE (top row), MICT (middle row) and IVL (bottom row) databases.

Table 5.1: Databases that contain JPEG distorted images.

Database	Images		Distortion	Methodology	Scale	Observers
LIVE [117]	29	768 x 512	about 8 levels	Single Stimulus	Continuous linear scale; Absolute category rating (5 categories)	20
MICT [107, 131]	14	768x512	6 levels	Single Stimulus	Absolute category rating (5 categories)	16
IVC [14, 15]	10	512x512	5 levels	Double Stimulus	impairment scale	15
TID [97]	25	512x384	4 levels	custom	custom	800
CSIQ [64]	30	512x512	4-5 levels	custom	Categorical	25
IVL	20	886x591	9 levels	Single Stimulus	Absolute category rating (3 categories)	31

Classification task

We focus for example on the **M3** metric and plot the logistic regression curve for the LIVE database in Figure 5.4. Two images, with different content and level of distortion, are highlighted for which the metric is not able to correctly predict the subjective scores.

In this chapter we address this IQA problem within a course to fine manner. That is, instead of obtaining a precise quality score we now focus on finding appropriate IQ classes. Once these classes have been identified, we can think about improving the IQA within each of the classes, where appropriate metrics for the specific applications can be applied. Therefore, we wonder if a direct classification obtained by thresholding the regression curve can achieve good performance.

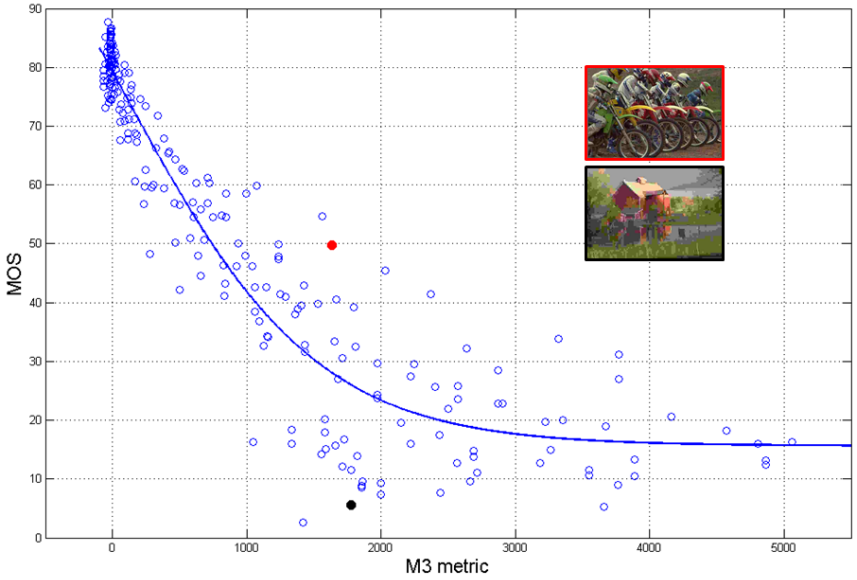

Figure 5.4: Logistic regression of the psychovisual data (MOS) and the **M3** NR metric [141] for the LIVE database. Two images with different content and different level of distortion but for which the metric values are similar are highlighted.

In Figure 5.5 (up), the MOS scores (of Figure 5.4) are grouped with respect to the five categorical attributes (*Bad*, *Poor*, *Fair*, *Good*, and *Excellent*) dividing linearly the continuous scale (1-100) in five intervals. These groups correspond to the ground truth of our classification problem. The predicted classes obtained thresholding directly the regression curve of Figure 5.4 are shown in Figure 5.5 (bottom). The four metrics' thresholds indicated in the figure are those corresponding to the regression function equal to MOS values of 20, 40, 60 and 80 respectively. The performance of this classification is reported in Table 5.2 in terms of confusion matrix. Each column of the matrix represents the instances in a predicted class, while each row represents the instances in an actual class. All correct predictions are located in the diagonal of the table. All the non-zero elements outside the diagonal represent misclassifications. The performance error is obtained as the ratio between the misclassified elements and the total number of images. Summarizing, for the **M3** metric and LIVE data, even if the Pearson correlation coefficient results equal to 0.95, the error performance for the classification task is significantly high and equal to 42%. Therefore, the classification framework applied to this particular single metric does not solve our initial problem. In the next sections the classification results for each of the NR metrics above listed are presented and compared with the results obtained using a combination of all these metrics.

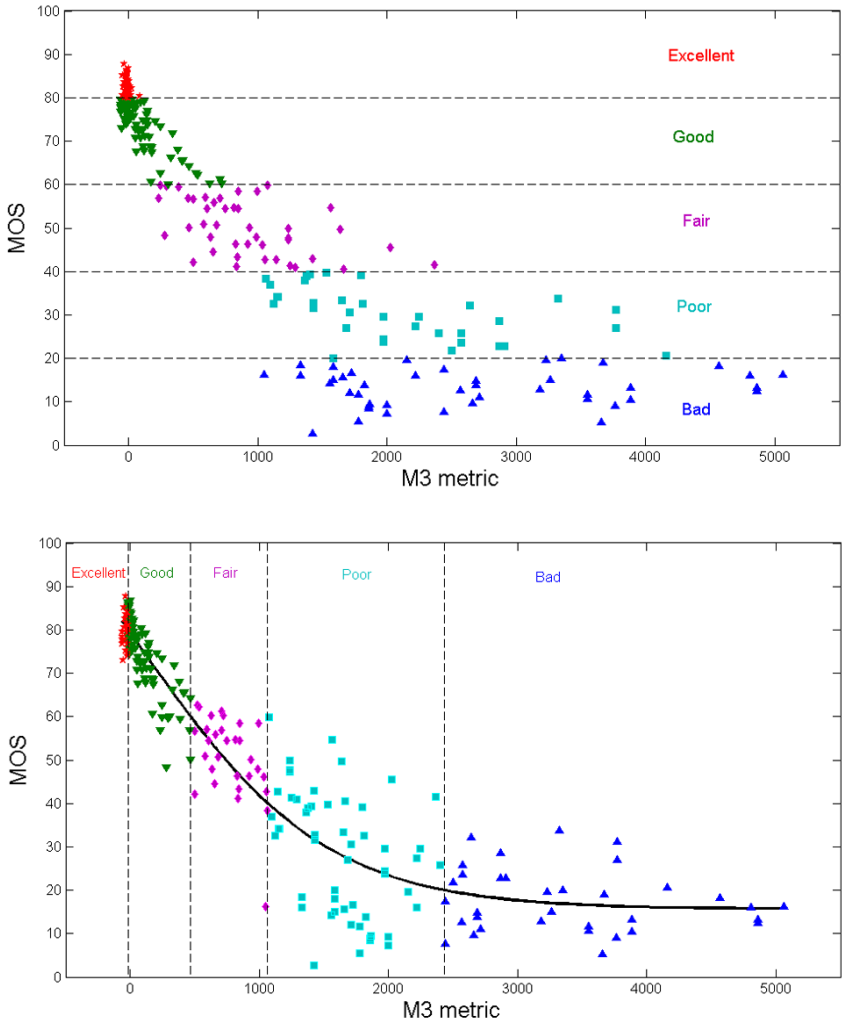

Figure 5.5: Up: MOS scores (of Figure 5.4) grouped with respect to the five categorical attributes. Bottom: the predicted classes obtained thresholding directly the regression curve of Figure 5.4.

5.3.1 Classification results on IVL, LIVE and MICT data

In this section we present and compare the performance of our classification scheme applied in different configurations in order to verify if a combining of metrics can achieve a better classification performance with respect to each single metric. We recall that LIVE data was obtained using an experimental set up showing 5 categorical attributes, while for the IVL experiment a continuous slider with 3 categorical classes (see Chapter 4) was used. Therefore with respect to the training set, we believe that the natural choice is the LIVE data for the case of 5 classes classifier and the IVL for the 3 classes classifier. As test sets we have considered LIVE, MICT, and IVL.

Table 5.3 reports the correspondences adopted in what follows between classes and categorical attributes.

Table 5.2: Confusion matrix for classification in five quality classes, obtained using the regression curve of **M3** metric and LIVE data.

class	*predicted*				
real	*Bad*	*Poor*	*Fair*	*Good*	*Excellent*
Bad	23	19	1	0	0
Poor	11	22	1	0	0
Fair	0	13	23	7	0
Good	0	0	5	52	11
Excellent	0	0	0	31	14

error = 42%

Table 5.3: Correspondences between classes and categorical attributes

5 classes classifier					
Categorical attributes	**Bad**	**Poor**	**Fair**	**Good**	**Excellent**
Classes	**1**	**2**	**3**	**4**	**5**
3 classes classifier					
Categorical attributes	**Bad**	**Poor**	**Fair**	**Good**	**Excellent**
Classes	**1**		**2**	**3**	

Initially we consider CART as machine learning approach, and the experiments performed are summarized in Table 5.4. To confirm our classification performances we also consider as machine learning method the widely used SVM. The first column of Table 5.4 indicates the label adopted for each of the listed configurations.

IQ classification in case of five classes

Let us compare the eleven configurations $C5_{Mi}$ (one for each of the eleven metrics, first row of Table 5.4), with the configuration that considers the combining of all the metrics $C5$ (second row of the same Table). In the training phase the whole LIVE database was used. As machine learning method we first consider CART, as it can provide a clear understanding on which features are the most significant within the classification task and we then confirm our results considering also the SVM approach (third row of the same Table). LIVE database was used in both training and test phases. As in general the classification trees obtained with CART are too large and thus tend to over-fit the data, we have pruned them back. The results were obtained applying cross-validation [9]. We have performed 29 rounds, partitioning the images of LIVE into 29 different couples of complementary subsets, to avoid data snooping. The splitting of the data was done carefully so that the image contents present in each training set did not appear in its test set. One image content is defined as all the distorted versions of a same original image. The performance of the classification is evaluated considering the validation results over the 29 rounds.

Table 5.4: Experimental configurations

Classifier label	Feature space	Machine learning method	Number of classes	Training set	Test set
$C5_{Mi}$ i=1,...,11	each of 11 NR metrics	CART	5	LIVE	LIVE
C_5	all 11 NR metrics	CART	5	LIVE	LIVE,IVL,MICT
S_5	all 11 NR metrics	SVM	5	LIVE	LIVE
$C3_{Mi}$ i=1,...,11	each of 11 NR metrics	CART	3	IVL	IVL
C_3	all 11 NR metrics	CART	3	IVL	LIVE,IVL,MICT
S_3	all 11 NR metrics	SVM	3	IVL	IVL

$C5_{Mi}$: training LIVE, test LIVE

As we observe from Table 5.5, the performance errors for the single-metric classifiers go from 27% ($C5_{M9}$), to a maximum value of 51.5%($C5_{M5}$). The misclassified classes are different for different metrics. $C5_{M2}$ and $C5_{M11}$ are not able to predict class 5 (i.e. the class corresponding to *Excellent* images) while $C5_{M3}$ is not able to predict class 2 (*Poor* images). In general, the greatest contribution to the errors come from misclassification of classes 4 and 5 (*Good* and *Excellent*). On the other hand, the classification of classes 1 and 2 (*Bad* and *Poor*) seem to be better achieved.

$C5$: training LIVE, test LIVE

The tree obtained for the classifier $C5$ on LIVE database is shown in Figure 5.6. We observe that two distortion specific metrics ($M3$ and $M4$) and two general purpose-metrics ($M8$ and $M9$) are the only ones taken into account for the classification task. Classes 4 and 5 are discriminated by using $M3$ and $M4$, while for classes 1, 2, and 3, metrics $M8$ and $M9$ have been chosen by the CART algorithm. With respect to these general purpose metrics, we note that they are OA-type and have been originally trained on the LIVE database. The performance error obtained with this tree results equal to 25.7% . Analyzing more in detail its confusion matrix (Table 5.6), we can note that:

- all the classes are predicted;
- misclassifications come only from the nearest classes;
- the greatest contribution to the overall error is due to the 26 images of real class 4, predicted in class 5.

Our initial hypothesis was to verify if the combination of metrics can improve the performances with respect to each of the single metrics. We note that it is partially confirmed in the present case of C5 on LIVE database: the $C5$ outperforms all the single classifiers $C5_{Mi}$. We note however that the improvement is not very noticeable for the case of $C5_{M9}$ (25.7% versus 27%). Recalling that metrics $M8$, $M9$, and $M10$ are OA metrics, trained on the same LIVE database, we wonder if the performance shown by $C5$ is strongly influenced by these metrics. Therefore, we have trained again the $C5$ but removing these three OA metrics from the feature space. The corresponding tree and confusion matrix are shown in

Figure 5.7 and Table 5.7 respectively. The performance error is now 28.3%, confirming that $M9$ which is an OA metric, is significant in the performance of the classifier. Also in this case we observe that a classifier that combines several metrics improves the performance of the classifiers corresponding to each single metric.

As already shown, in the case of the single classifiers $C5_{Mi}$ (with i from 1 to 11), there exist another way to achieve the classification task. We could have started from the continuous quality scores of each of the single NR metrics and threshold their regression curves. We have verified that both approaches (quantization of the regression curves and direct classification) perform similar. However, no such equivalent method can be thought for the case of $C5$ that combines the pool of eleven metrics.

Table 5.5: Five classes: Confusion matrices for CART classification trained and tested on LIVE, using each metric as a single feature ($M1 - M11$).

$C5_{M1}$

class	predicted				
real	**1**	**2**	**3**	**4**	**5**
1	40	3	0	0	0
2	4	23	7	0	0
3	0	1	30	12	0
4	0	0	3	34	31
5	0	0	0	11	34

error = 30.9%

$C5_{M2}$

class	predicted				
real	**1**	**2**	**3**	**4**	**5**
1	39	4	0	0	0
2	11	13	9	1	0
3	0	7	21	15	0
4	0	3	7	58	0
5	0	0	3	42	0

error = 43.7%

$C5_{M3}$

class	predicted				
real	**1**	**2**	**3**	**4**	**5**
1	42	0	1	0	0
2	29	0	5	0	0
3	7	0	25	11	0
4	0	0	8	33	27
5	0	0	0	5	40

error = 39.9%

$C5_{M4}$

class	predicted				
real	**1**	**2**	**3**	**4**	**5**
1	38	5	0	0	0
2	13	9	12	0	0
3	0	6	31	6	0
4	0	0	6	29	33
5	0	0	0	10	35

error = 39%

$C5_{M5}$

class	predicted				
real	**1**	**2**	**3**	**4**	**5**
1	37	4	2	0	0
2	13	2	17	2	0
3	4	3	21	15	0
4	0	0	14	50	4
5	0	2	3	37	3

error = 51.1%

$C5_{M6}$

class	predicted				
real	**1**	**2**	**3**	**4**	**5**
1	40	3	0	0	0
2	6	13	14	1	0
3	0	3	25	14	1
4	0	0	9	37	22
5	0	0	1	21	23

error = 40.7%

$C5_{M7}$

class	predicted				
real	**1**	**2**	**3**	**4**	**5**
1	40	3	0	0	0
2	6	14	13	1	0
3	0	8	27	8	0
4	0	0	13	41	20
5	0	0	0	22	23

error = 37.7%

$C5_{M8}$

class	predicted				
real	**1**	**2**	**3**	**4**	**5**
1	37	6	0	0	0
2	8	22	4	0	0
3	0	3	35	5	0
4	0	0	12	27	29
5	0	0	0	23	22

error = 38.6%

$C5_{M9}$

class	predicted				
real	**1**	**2**	**3**	**4**	**5**
1	40	3	0	0	0
2	3	26	5	0	0
3	0	4	31	8	0
4	0	0	8	49	11
5	0	0	0	21	24

error = 27.0%

$C5_{M10}$

class	predicted				
real	**1**	**2**	**3**	**4**	**5**
1	38	4	1	0	0
2	18	7	9	0	0
3	4	6	23	10	0
4	0	0	10	42	16
5	0	0	0	15	30

error = 39.9%

$C5_{M11}$

class	predicted				
real	**1**	**2**	**3**	**4**	**5**
1	35	8	0	0	0
2	8	13	12	1	0
3	0	2	25	16	0
4	0	0	6	62	0
5	0	0	0	45	0

error =42.0 %

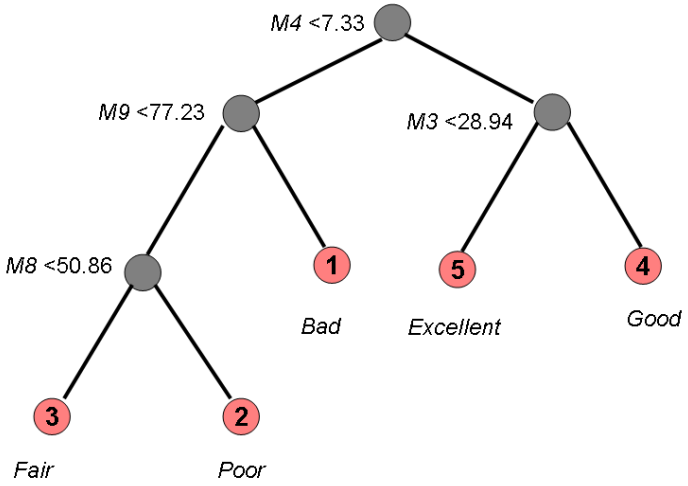

Figure 5.6: CART classifier $C5$ obtained considering the eleven metrics.

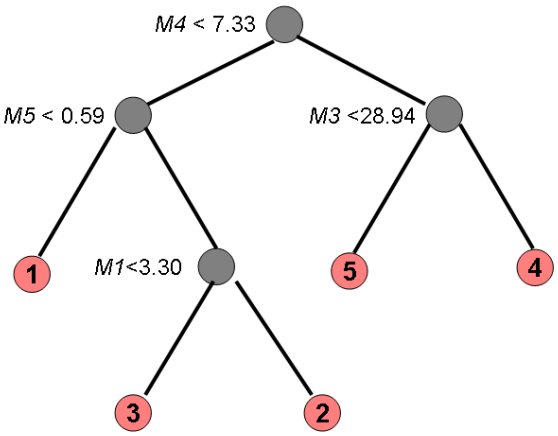

Figure 5.7: CART classifier obtained considering only OU metrics and metrics not trained on the LIVE data.

Table 5.6: Five classes: Confusion matrix for CART $C5$ classification trained and tested on LIVE, using all the eleven metrics as feature space.

class	*predicted*				
real	**1**	**2**	**3**	**4**	**5**
1	40	3	0	0	0
2	3	26	5	0	0
3	0	6	29	8	0
4	0	0	5	37	26
5	0	0	0	4	41

error = 25.7%

Table 5.7: Five classes: Confusion matrix for CART classification trained and tested on LIVE, using, as feature space, only metrics that are Opinion Unaware (UA) or not trained on the LIVE data (that is the first seven metrics and the eleventh).

class	*predicted*				
real	**1**	**2**	**3**	**4**	**5**
1	37	6	0	0	0
2	5	21	8	0	0
3	0	3	34	6	0
4	0	0	6	36	26
5	0	0	0	6	39

error = 28.3%

$C5$: training LIVE, test MICT and IVL

Finally, we test the classifier $C5$ on the MICT and IVL databases. Let us recall that the distortion distributions of LIVE, MICT, and IVL are significantly different, specially in the range of high compression levels. Therefore, the performance of $C5$ trained on LIVE, and tested on MICT and IVL are very low (errors greater than 50%). To better understand how the misclassifications are distributed, in Figure 5.8 the histograms of the subjectivel classes (left) and classes predicted with $C5$ (right) are compared for both MICT (top) and IVL (bottom) databases. The missed estimated classes (1 and 2) for both MICT and IVL correspond to severe level of distortions that are present in the training set (LIVE) but not in the test sets.

$S5$: training LIVE, test LIVE

We now apply SVM when the feature space is composed of the eleven metrics (see third row of Table 5.4). LIVE is used for both training and test phases. When training an SVM, a very important step is the choice of the kernel function and the setting of the corresponding parameters. We have followed the proposal of Hsu et al. [48] and after a scaling of the data, the Radial Basis Function (RBF) kernel was chosen. The penalty term and the parameter of the RBF used in the training and testing phase are found using a cross-validation procedure.

We present in Table 5.8 the confusion matrix corresponding to $S5$. As it can be seen from the Table, again the greatest contribution to the misclassification comes from classes 4 and 5 as it occured for $C5$. The performance error (24.5%) is also similar to the one obtained for $C5$ (25.7%).

Table 5.8: Five classes: Confusion matrix for SVM classifier $S5$, trained and tested on LIVE, using all the eleven metrics as feature space.

class	*predicted*				
real	**1**	**2**	**3**	**4**	**5**
1	39	4	0	0	0
2	5	26	3	0	0
3	0	3	34	6	0
4	0	0	5	45	18
5	0	0	0	13	32

error = 24.5%

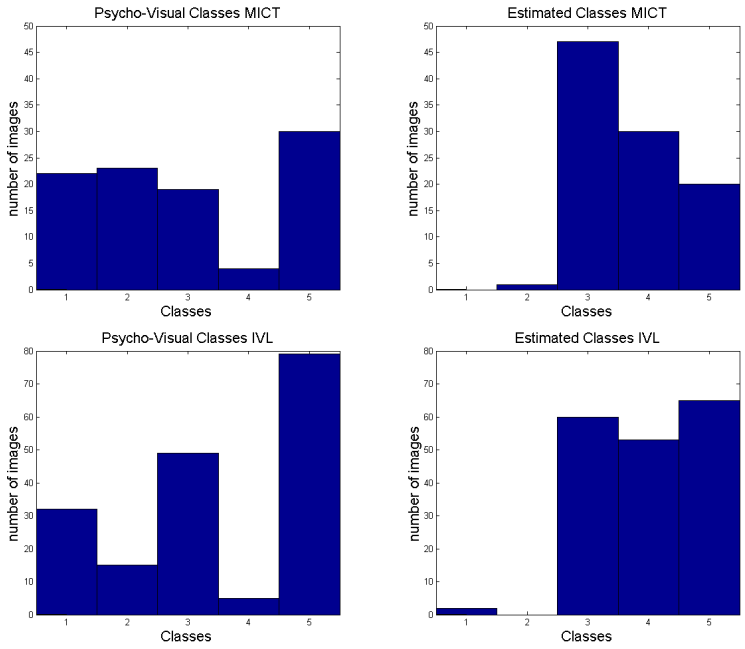

Figure 5.8: Histograms of the real classes (left) and predicted classes (right) for both MICT (top) and IVL (bottom) databases, using the $C5$ tree trained on the LIVE dataset.

IQ classification in case of three classes

In this section, we compare the eleven configurations $C3_{Mi}$ (corresponding to each of the NR metrics, see Table 5.4), with the configuration that considers the combining of all the metrics $C3$. We first consider CART as machine learning method and to confirm our results we then use the SVM approach to obtain the $S3$ classifier (see Table 5.4). We have trained

all the classifiers using all the 180 images of the IVL database, and we have evaluated their performances on the same database, using cross-validation. We have performed 20 rounds, partitioning the 180 images into 20 different couples of complementary subsets of 171 and 9 images respectively, to avoid data snooping. For each round, the training phase was performed on the subset with 171 images and the test on the other subset. The splitting of the data was done carefully so that the image contents present in each training set did not appear in its test set. One image content is defined as all the distorted versions of a same original image. The performance of the classification is evaluated considering the validation results over the 20 rounds. As in general the classification trees are too large and thus tend to over-fit the data, we have pruned them back.

The correspondences between the three classes and the categorical attributes are those indicated in Table 5.3.

$C3_{Mi}$: training IVL, test IVL

In Table 5.9 the results for the eleven classifiers $C3_{Mi}$, when IVL is used as test set, are reported in terms of confusion matrices and error performances. Most of the single classifiers misclassify either class 1 into 3 and/or class 3 into 1. Classifiers C_{M2}, C_{M5}, C_{M8} and C_{M9} are not able to predict class 2.

$C3$: training IVL, test IVL, LIVE and MICT

We report in Table 5.10 the results for $C3$ classifier when the IVL (with cross validation), LIVE, and MICT databases are used as test sets respectively. The pruned tree corresponding to the $C3$ classifier is reported in Figure 5.9.

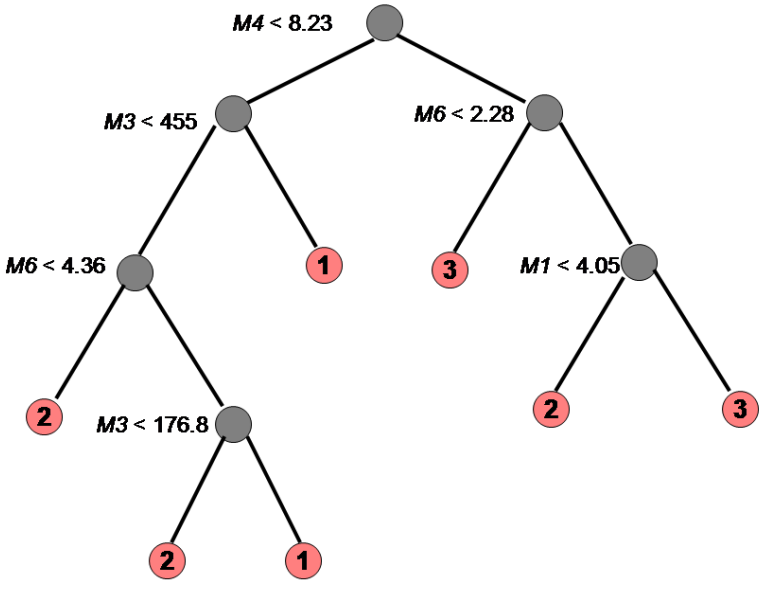

Figure 5.9: CART classifier $C3$ obtained considering the eleven metrics and trained on the IVL database

Table 5.9: Three classes: Confusion matrices for CART classification trained and tested on IVL, using each metric as a single feature ($M1 - M11$).

$C3_{M1}$

class	*predicted*		
real	**1**	**2**	**3**
1	35	14	0
2	18	16	8
3	0	5	84

error = 25%

$C3_{M2}$

class	*predicted*		
real	**1**	**2**	**3**
1	24	0	25
2	9	0	33
3	4	0	85

error = 39.4%

$C3_{M3}$

class	*predicted*		
real	**1**	**2**	**3**
1	31	1	1
2	7	23	12
3	0	4	85

error = 22.8%

$C3_{M4}$

class	*predicted*		
real	**1**	**2**	**3**
1	31	18	0
2	11	21	10
3	1	4	84

error = 24.4%

$C3_{M5}$

class	*predicted*		
real	**1**	**2**	**3**
1	44	1	4
2	26	0	16
3	23	0	66

error = 38.9%

$C3_{M6}$

class	*predicted*		
real	**1**	**2**	**3**
1	40	9	0
2	10	25	7
3	2	4	83

error = 17.8%

$C3_{M7}$

class	*predicted*		
real	**1**	**2**	**3**
1	36	13	0
2	21	12	9
3	5	5	79

error = 29.4%

$C3_{M8}$

class	*predicted*		
real	**1**	**2**	**3**
1	24	0	25
2	10	0	32
3	8	0	81

error = 41.7%

$C3_{M9}$

class	*predicted*		
real	**1**	**2**	**3**
1	37	0	12
2	18	0	24
3	18	0	71

error = 40.0%

$C3_{M10}$

class	*predicted*		
real	**1**	**2**	**3**
1	37	7	5
2	20	11	11
3	2	4	83

error = 27.2%

$C3_{M11}$

class	*predicted*		
real	**1**	**2**	**3**
1	10	7	32
2	12	1	29
3	19	11	59

error =61.1 %

From Tables 5.9 and 5.10 we observe that the errors of all $C3_{Mi}$ result greater than the corresponding one to $C3$ (13.3%). The metrics $M1$, $M3$, $M4$ and $M6$ have been chosen by the algorithm to construct the tree (Figure 5.9). We note that these metrics are the ones that present the lowest errors when used individually to construct the classifier. In order to further compare the real and predicted classes by $C3$, the corresponding bar diagrams as a function of the distortion level ($bppR$) are plot in Figure 5.10 for the IVL database. Each column sums up to 20 (the number of reference images) and reports the proportion of the predicted classes (on the right of the figure) and the psycho-visual data (on the left of the figure).

Table 5.10: Three classes: Confusion matrices for $C3$ tested on IVL, LIVE and MICT databases

IVL

class	*predicted*		
real	**1**	**2**	**3**
1	41	8	0
2	5	30	7
3	0	4	85

error = 13.3%

LIVE

class	*predicted*		
real	**1**	**2**	**3**
1	77	0	10
2	40	3	0
3	10	16	87

error = 28.3%

MICT

class	*predicted*		
real	**1**	**2**	**3**
1	43	2	0
2	8	10	1
3	0	9	25

error = 20.4%

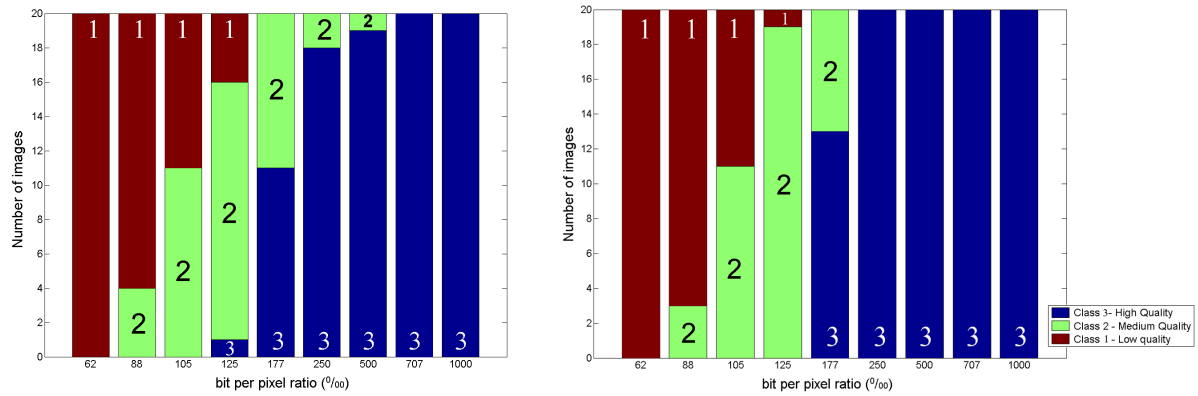

Figure 5.10: Comparison of the bar diagram of the classes assigned by the psycho-visual experiment, on the left (Figure 4.5), with the bar diagram obtained applying $C3$, on the right.

The classification error obtained applying $C3$ to the LIVE database is 28.3%. As it is observed from the asymmetry of its confusion matrix, the error is mainly due to the misclassification of images judged *Good* and *Excellent* (third row) or *Fair* (second row) that the classifier assigned in the worst class. This fact can be attributed to the different distribution of distortion levels between IVL (that has trained the classifier) and LIVE, as shown in Figure 5.2. The results of applying $C3$ on the MICT data results in a classification error of 20.4%. The better performance (with respect to LIVE results) can be again attributed to the more similar distribution of distortion levels between MICT and IVL databases (Figure 5.2).

Weighted $C3$: training IVL, test IVL

In a specific application domain, like the printing task mentioned in the Introduction, misclassified classes may have different weights. In such printing task we may prefer to misclassify images of class 1 or class 3 into class 2 (that corresponds to the class of images that require a user intervention before processing) instead of misclassifying images of class 2 (i.e. images to be evaluated) into class 1 or 3 (images definitively rejected or accepted). Let us note that with our $C3$ classifier none of the images of actual class 1 are predicted as class 3 while none of the images of actual class 3 are predicted as class 1 (see corresponding

confusion matrix of Table 5.10). We have thus trained another tree, assigning different misclassification weights. In particular, using the misclassification weights reported in Table 5.11, we have obtained the tree shown in Figure 5.11. The confusion matrix is reported in Table 5.11. The performance error, which is now 14.4% results greater than in the previous case, but less images of class 2 were misclassified.

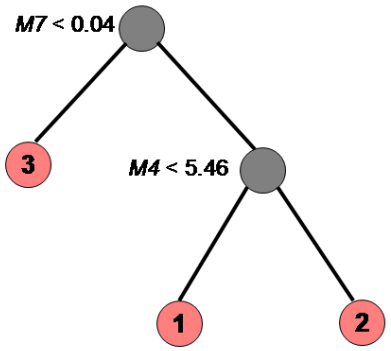

Figure 5.11: Classification tree with different misclassification weights.

Table 5.11: $C3$ tested on IVL, with misclassification weights

Confusion matrix

class	*predicted*		
real	**1**	**2**	**3**
1	37	12	0
2	3	36	3
3	0	8	81

error = 14.4%

misclassification weights

	predicted classes		
real classes	**1**	**2**	**3**
1	0	1	1
2	2	0	2
3	1	1	0

$S3$: IVL training, IVL test

Finally, we consider the SVM classifier $S3$ trained on the whole IVL dataset. After a scaling of the data, the Radial Basis Function (RBF) kernel was chosen. The penalty term and the parameter of the RBF used in the training and testing phase are found using a cross-validation procedure. As we have noted before for $S5$, the common approach of training

the classifier with for instance 2/3 of the whole dataset and testing it on the remaining 1/3 could be not correct as it could strongly depend on the training set used. For this reason we have adopted the same cross validation approach described for CART. Table 5.12 reports the classification results obtained for $S3$ when IVL, LIVE, and MICT are used as test sets respectively.

Table 5.12: Confusion matrices for $S3$ tested on IVL, LIVE, and MICT databases

IVL

class	*predicted*		
real	**1**	**2**	**3**
1	44	5	0
2	7	28	7
3	0	5	84

error = 13.3%

LIVE

class	*predicted*		
real	**1**	**2**	**3**
1	77	0	0
2	38	5	0
3	10	30	73

error = 37.3%

MICT

class	*predicted*		
real	**1**	**2**	**3**
1	41	1	0
2	6	13	0
3	0	13	21

error = 23.5%

Summarizing, from the comparison of the results obtained for 5 classes versus 3 classes, we observe that $C3$ (training and test IVL) achieves better the classification task than $C5$ (training and test LIVE). The better correlation of $C3$ with the corresponding subjective data, compared to $C5$, could be attributed to the easier subjective task in the case of 3 classes. In fact, from the observer point of view it seems simpler to judge the quality of an image within one of 3 classes (*Excellent*, *Fair* or *Bad*) than making a finer analysis to classify it within 5 classes (Is the image excellent or good? Is the image poor or bad?). That is, as observers we are able to decide almost instantly whether a particular image is of good or poor quality but for us to quantify how good an image is, and the scale of quality to be used is far more difficult.

5.4 Noise IQA classification on IVL data

In this section we aim to classify the noisy images of the IVL database within the three groups of high, medium and low quality. We choose the metric by Immerkaer [51] because of its simplicity and the high performance shown for the case of additive Gaussian noise (the case of our IVL noise database). We consider two general purpose metrics [4, 78] as well.

We choose here to apply a direct classification task by thresholding separately each of the regression curves obtained. The real and predicted quality classes are obtained in a similar way as shown Figure 5.5 but in this case we are interested only in three classes. The performance of the classification task is reported in Table 5.13 in terms of confusion matrix and for each of the NR metrics here considered. We note that the best classification performance is obtained with Immerkaer metric [51] and as desired, no missclassification between classes 1 and 3 is observed. In Figure 5.12 the thresholds predicted by this metric for the classification of the noise IVL data are shown.

Table 5.13: Confusion matrices for 3 classes classifiers corresponding to each of the NR metrics here considered and the noise IVL database.

Immerkaer

class	*predicted*		
real	**1**	**2**	**3**
1	65	5	0
2	4	58	0
3	0	14	54

error = 11.5%

BRISQUE

class	*predicted*		
real	**1**	**2**	**3**
1	66	4	0
2	6	48	8
3	0	13	55

error = 15.5%

NIQE

class	*predicted*		
real	**1**	**2**	**3**
1	39	31	0
2	8	46	8
3	3	35	30

error = 42.5%

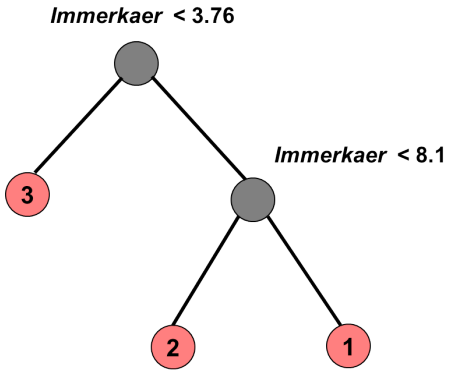

Figure 5.12: Classification thresholds predicted by Immerkaer metric when applied to noise IVL data.

5.5 Conclusions

In the present Chapter we have approached the NR-IQA field by focusing on a direct classification problem that combines different objective metrics. To this end we have proposed a framework based on a machine learning classification, where NR metrics are considered as features and the assigned classes are obtained from the psychovisual data. Eleven NR metrics have been considered for the JPEG-blockiness case: seven specific ones and four general purpose. In order to confirm our initial hypothesis, that is the combining of metrics can outperform a single method, we have considered different feature spaces: each metric individually and the pool of the eleven metrics together.

Classification within five and three classes was addressed. In the former case, the five classes are in correspondence to the five categories recommended by the ITU [2] (excellent, good, fair, poor, and bad) when designing image quality experiments. In the later case we were interested in classifying images as high, medium or low quality ones. For the training of the five classes classifier ($C5$ and $S5$), the well known LIVE database was used. It was then tested on LIVE, IVL and MICT databases. For the three classes classifier ($C3$ and $S3$), the database IVL was properly generated for this scope and psycho-visual experiments were conducted to identify the high, medium and low quality images. The IVL dataset was used for the training phase while for testing we used IVL, LIVE and MICT databases.

Considering the pool of all the metrics as feature space, the misclassification error ob-

tained in the testing phase when CART approach was used, was about 25% for $C5$ (on LIVE) and 13% for $C3$ (on IVL). The greater misclassification errors obtained when tested on different databases may be partially attributed to the different distribution of distorted images within the considered databases, confirming that datasets coming from different psycho-visual experimental setups should be carefully compared. With respect to the SVM approach, similar performances were obtained for $S5$ and $S3$.

We have also applied the CART methodology using each NR metric as single feature. Our initial hypothesis was confirmed for both $C5$ and $C3$: the classifiers that combine all the metrics improve the performances of each of the eleven single classifiers (corresponding to the single metrics taken into account individually).

The presented classification scheme could be useful within an image quality control workflow chain. For example, if the final goal is printing the images, the best ones can be directly printed while the worst ones can be automatically discarded. The images belonging to medium quality class should be eventually evaluated manually by an operator. In order to help the operator, in a previous work, we have proposed an interactive tool [25], that permits to apply several NR metrics, not only globally but also locally.

In order to improve our results, it could be useful to integrate the classification scheme with a regression module. For example Marini et al. [76] use regression trees to combine quality metrics and texture/structure descriptors, based on wavelets, focusing on JPEG artifacts. The idea is to find "content weighting factors" that take into account the influence of image details in the perception of the distortions.

Chapter 6

Improving regression between subjective and objective data

It is in general assumed that subjective methods produce an actual estimate of the perceived quality while objective methods produce values that should be correlated with human perceptions as best as possible. As already reviewed in Chapter 2, it is customary to apply a nonlinear transformation to the predicted scores so as to bring the predictions on the same scale as the subjective data and to attempt to obtain a linear relationship between the predictions and the opinion scores. The VQEG [136] suggests the use of logistic or polynomial functions. Recently, it has also been proposed a Monotonic Regression (MR) [44].

In this chapter we aim to improve the agreement between NR metrics and subjective data. Different hypothesis will be tested. The first and second ones consider weighting the NR metrics by factors depending on the saliency maps and on the spectral frequency of the images respectively. The third hypothesis is based on grouping the images according to their spatial complexity. Each strategy is investigated and tested for the case of JPEG-blockiness artifacts.

6.1 Strategy based on saliency maps

Research in image quality assessment attempts to further improve the reliability of objective metrics by taking into account the human visual system characteristics. One of such features is the visual attention mechanism that is responsible for defining which areas of the scene are relevant and should be attended. There are two visual selection mechanisms: bottom-up and top-down. The bottom-up mechanism is an automated selection that is controlled mostly by the signal. It is fast and short lasting, being performed as a response to low-level features that are perceived as visually salient. The top-down mechanism is controlled by higher cognitive factors and external influences, such as semantic information, viewing task, and personal preferences, context. It is slower and requires a voluntary effort.

A recent development in the area of image quality consists of trying to incorporate aspects of visual attention in the design of visual quality metrics, mostly using the assumption that visual distortions appearing in less salient areas might be less visible and, therefore, less annoying. Saliency maps, that are bottom-up models of attention, are frequently considered within the IQA field during the last years. In the case of FR metrics, the typical integration strategy consists of the multiplication of each local objective metric value with the corresponding saliency map. However, the results obtained by different groups are not yet

conclusive [133]. 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 [65, 128, 75, 74, 82] showed that better agreement with subjective scores can be produced for IQA metrics when saliency maps are taken into account in the metrics' evaluation. On the other hand, others claim that for example, MSE and SSIM do not show a clear improvement [86, 156]. In particular, the results from [86] 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. Among the works that have reported some improvement, most use subjective saliency maps, i.e. saliency maps generated from eye-tracking data obtained experimentally [70]. Although subjective saliency maps are considered as the ground-truth in visual attention, they cannot be used in real-time applications. Therefore, computational models of visual attention have to be considered [55].

With respect to the integration of saliency maps on RR or NR methods, less research has been done up to date. Since, in general, the NR metrics do not generate an "error" map as a final step, the inclusion of saliency maps complies with the local computation strategy involved in each of the specific NR metrics. In the present section we propose to weight the input image by different saliency maps and then simply apply the JPEG-blockiness metrics to these weighted-input images.

Four popular bottom-up visual attention computational models are taken into account: Itti and Koch [55] (hereafter named $S1$), Achanta et al. [6] (named $S2$), Harel et al. [45] (named $S3$) and Hou and Zhang [47] (named $S4$). For a given image, these models generate a gray-scale saliency map indicating image regions that are most likely to attract attention. In the saliency maps, higher luminance values correspond to higher saliency pixels, while lower values correspond to lower saliency ones. Experiments have been performed on JPEG LIVE data. For each of the distorted images, the four saliency maps have been evaluated. We have also considered the saliency maps generated from eye-tracking data (named $S5$) obtained on each of the 29 original images of the database [70]. In Figure 6.1 an original image from LIVE database, the four computational models of saliency and the map based on the eye-tracking data are shown.

Figure 6.1: Saliency maps. First row: an original image from LIVE and S1, second row: S2 and S3, third row: S4 and S5.

Assuming that the saliency map $sal(x, y)$ is normalized in the interval [0,1], the complement of the saliency $[1 - sal(x, y)]$ is also taken into account within the present strategy. The reference image with the corresponding complement saliency maps are depicted in Figure 6.2.

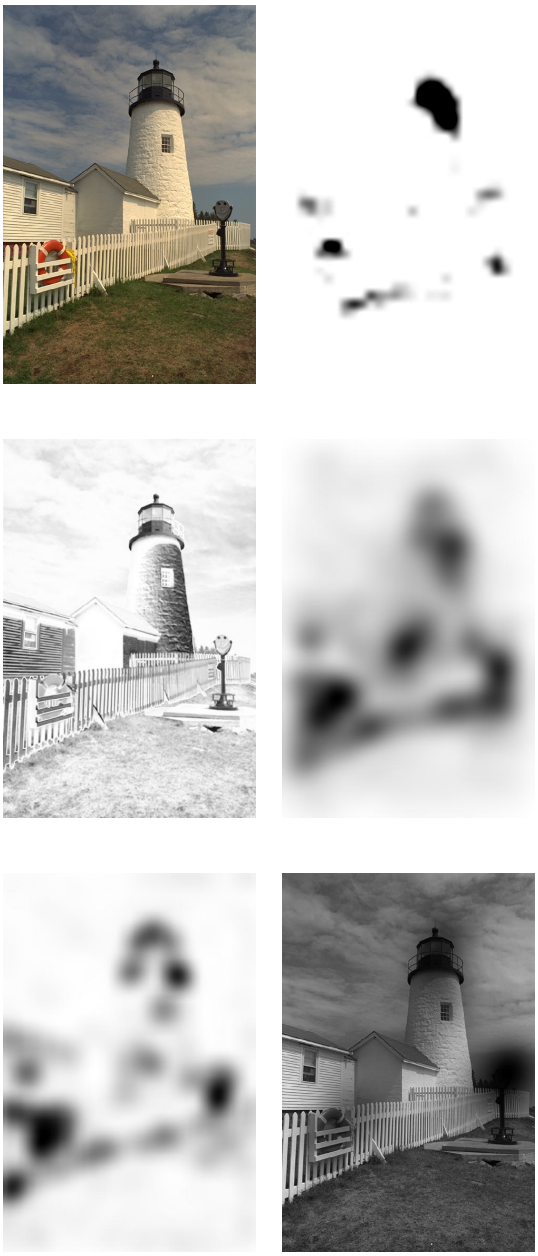

Figure 6.2: Complement of the saliency maps. First row: original image and (1-S1), second row: (1-S2) and (1-S3), third row: (1-S4) and (1-S5).

Each of the saliency maps are then binarized, where a weight equal to one is applied to the salient regions and 0.5 otherwise. In the case of the complement saliency maps, the weights are 0.5 for the salient region and 1 otherwise. The threshold value for the binarization process is obtained using the Otsu method [93]. In Figure 6.3 a reference image from LIVE together with the image multiplied by the binarized saliency map and the binarized complement of

the map are shown for the computational model of Harel et al. [45].

Figure 6.3: Original image (left), image weighted by the binarized map S3 (center), image weighted by the binarized complementary map (right).

After weighting the input distorted image by its binarized saliency map (or its binarized complementary map), six specific JPEG-blockiness metrics have been evaluated: Pan [94], WSB [142], WBE [141], GBIM [152], Muijs and Kirenko [83] Chen and Bloom [21]; as well as three general purpose ones: BIQI [81], BRISQUE [4] and NIQE [78].

In Table 6.1 the performance of the NR methods is evaluated in terms of the PCC coefficient after logistic regression with the subjective scores of the LIVE database. The first row of the Table reports the coefficient values without saliency and each of the following rows correspond to the different computational models ($S1$ to $S4$) and the eye-tracking data ($S5$) as well as the complement of these saliency maps $S1C$ to $S5C$).

PCC	GBIM	PAN	WBE	WSB	MUIJS	CHEN	BIQI	BRISQUE	NIQE
No sal.	0,9589	0,8893	0,9170	0,9787	0,9295	0,9450	0,9631	0,9874	0,9518
S1	**0,9619**	0,8325	0,9085	0,9704	0,7218	0,9210	0,9482	0,9703	0,9368
S1C	**0,9605**	0,8760	0,9129	0,9785	0,9248	0,9368	**0,9645**	0,9831	0,9493
S2	0,9210	0,8108	0,9151	0,9749	0,9063	0,9137	0,8720	0,9710	0,9077
S2C	**0,9600**	0,8518	**0,9193**	0,9776	0,7262	0,9218	0,9344	0,9742	0,9291
S3	**0,9603**	0,8356	0,9121	0,9749	0,9142	0,9236	0,9514	0,9750	0,9409
S3C	0,9292	0,8677	0,9086	0,9732	0,9195	0,9315	**0,9659**	0,9807	0,9418
S4	**0,9628**	0,8318	0,9119	0,9760	0,9122	0,9212	0,9539	0,9745	0,9374
S4C	**0,9605**	0,8760	0,9129	0,9785	0,9248	0,9368	**0,9645**	0,9831	0,9493
S5	0,9551	0,8398	0,9145	0,9732	0,9141	0,9235	0,9561	0,9751	0,9409
S5C	0,9405	0,8727	0,9087	0,9749	0,9202	0,9337	**0,9684**	0,9834	0,9480

Table 6.1: PCC for NR metrics when the distorted images are weighted by saliency maps.

From Table 6.1 we observe that few combinations of metrics and saliency maps show improvements with respect to the traditional method (without saliency). The metric GBIM seem to be the one that better benefits from the inclusion of the saliency maps. Also BIQI show some improvement for nearly all the complementary maps. We can say that in general no better performances are obtained with the saliency-based strategy. An important aspect is the integration strategy of the saliency map within the metric. We have here proposed a

very simple one but it should be crucial to analyze different integration strategies to include saliency maps within the NR metrics [57].

6.2 Strategy based on the spectral frequency analysis

This strategy is based on the analysis of the spectral frequency of the images. To this aim, we compute the frequency $\overline{\rho}$ in the Fourier domain, corresponding to the 99% of the image energy. Starting from the image spectrum in polar coordinates $S(\rho, \theta)$ we have that:

$$E = \int_0^{\infty} \int_{\theta=0}^{2\pi} |S(\rho, \theta)|^2 d\rho d\theta \tag{6.1}$$

and we define $\overline{\rho}$ as follows:

$$0.99E = \int_0^{\overline{\rho}} \int_{\theta=0}^{2\pi} |S(\rho, \theta)|^2 d\rho d\theta \tag{6.2}$$

Each of the single NR metrics is weighted by a frequency dependent factor f_n obtained normalizing the frequency $\overline{\rho}$ with respect to the NyQuist frequency ρ_{max}

$$f_n = \frac{\overline{\rho}}{2\rho max} \tag{6.3}$$

In Figure 6.4 the original images of the LIVE database are sorted with respect to increasing frequency, starting from the top left corner, to the bottom right one.

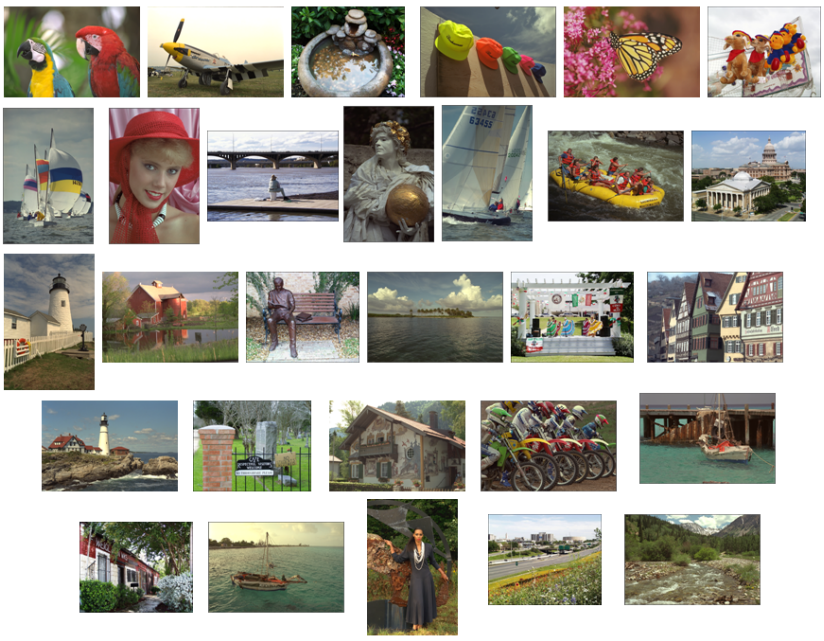

Figure 6.4: The 29 reference images of the LIVE database sorted with respect to increasing frequency, starting from the top left corner, to the bottom right one.

For the present analysis we consider six NR JPEG-blockiness metrics: Pan et al. [94], Vlachos [135], WSB [142], WBE [141], GBIM [152], Chen and Bloom [21] and three general purpose blind metrics: BIQI [81], BRISQUE [4] and BLIINDS [104].

The regression between the subjective scores and the NR metrics with and without the frequency-weighting factor are compared in Tables 6.2 and 6.3 for LIVE and CSIQ databases respectively.

Metric		*Freq-Weight*
	PCC	PCC(Freq-Weight)
PAN et al. [94]	0.8599	**0.8628**
VLACHOS [135]	0.8307	**0.8534**
WSB [142]	0.9414	**0.9438**
WBE [141]	0.8304	**0.9356**
GBIM [152]	0.9485	**0.9523**
CHEN and BLOOM [21]	0.9447	**0.9459**
BIQI [81]	0.9181	0.9064
BRISQUE [4]	0.9345	0.9329
BLIINDS [104]	0.9105	**0.9127**

Table 6.2: Comparison of the PCC corresponding to each metric and its frequency-weighted version for the LIVE database.

Metric		*Freq-Weight*
	PCC	PCC(Freq-Weight)
PAN et al. [94]	0.8739	**0.8917**
VLACHOS [135]	0.9063	0.8693
WSB	0.9483	0.9412
WBE [142]	0.9105	0.9091
WSB [141]	0.9348	**0.9429**
GBIM [152]	0.9186	**0.9267**
BIQI [81]	0.8475	0.8438
BRISQUE[4]	0.9086	**0.9149**
BLIINDS [104]	0.8926	**0.9193**

Table 6.3: Comparison of the PCC corresponding to each metric and its frequency-weighted version for the CSIQ database.

In the tables, the bold characters indicate the best performance achieved for each of the single metrics. For the LIVE database, the frequency strategy improve the results for seven out of the nine metrics considered: improvements are observed for all the specific JPEG-blockiness metrics and for only one general purpose method. For the CSIQ database, improvements are observed for five out of nine NR metrics: three JPEG specific and two general purpose.

6.3 Strategy based on the image complexity

Our working hypothesis is that regression can be improved if performed within a group of images that present similar contents in terms of low level features (not semantic content).

The complexity of an image tells many aspects of the image content. Therefore, the criteria we choose to divide the images in different groups is the image complexity.

The effect of content dependency on objective image quality metrics has been already addressed in the literature. For example, the authors in [132] have addressed the problem of scene dependency and scene susceptibility in image quality assessments and have proposed image analysis as a means to group test scenes, according to basic inherent scene properties that human observers refer to when they judge the quality of images. Experimental work has been carried out for JPEG and JPEG2000 distortions. Oh et al. [88] have analyzed the degree of correlation between scene descriptors (first and second order statistical measurements) and scene susceptibility parameters for noisiness and sharpness. Using different scene descriptors and applying K-mean clustering, three groups of scenes were successfully derived, depending on the relationship between the susceptibility to sharpness and noisiness distortions.

6.3.1 Computing image complexity

There exists no unique definition of the complexity of an image. Researchers from various fields have proposed different measures to estimate image complexity. Fuzzy approaches [16], information-theoretical based techniques [101] and independent component analysis [96] have been proposed in the literature to determine the complexity of an image. Recently, Yu and Winkler [157] have explored objective measures of complexity that are based on compression. They have shown that spatial information (SI) measures strongly correlate with compression-based complexity measures. Among the commonly used SI measures, the mean of the edge magnitude is shown to be the best predictor. Oliva et al. [89] performed an experimental session to study the representation of visual complexity for real-world scene images. The results obtained are consistent with a multi-dimensional representation of visual complexity (quantity of objects, clutter, openness, symmetry, organization, variety of colors).

In the present work we have performed a psycho-visual experiment: the complexity of each of the 29 original LIVE images has been judged by four members of our laboratory. The observers have been asked to classify each of the images in the following groups: high, medium or low complexity images. The result obtained from the grouping task is shown in Figure 6.5. Even if the present classification results can be considered as preliminary because of the small number of participants involved in the experiments, the results and conclusions achieved in this section are still valid and do not depend on slight variations of the subjective grouping.

We are interested in finding the low level image features that determine the complexity level of the images. To this end, for each of the 29 reference images of the LIVE database we have evaluated first (average, standard deviation, smoothness, skewness, entropy) and second order statistics (contrast, correlation, energy, spectral texture, coarseness, directionality). Also the complexity index proposed by Chacon et al. [16] has been taken into account. This method determines the complexity of an image based on the analysis of its edge level percentages. Using CART method [13], the feature space was built using the above mentioned statistics plus the complexity index. For the training of the classifier, the classes indicated in Figure 6.5 have been used. A classification tree is obtained where only the complexity index is the selected feature that discriminates among the three complexity classes. This tree is then used to classify the distorted images. As an example, the references images of the CSIQ and MICT databases are grouped as reported in Figures 6.6 and 6.7 respectively. In general, the complexity classes assigned to the distorted images are the same as those corresponding to the original one. In the few cases where this is not true, the complexity of the distorted image results in the following class: if it was originally classified

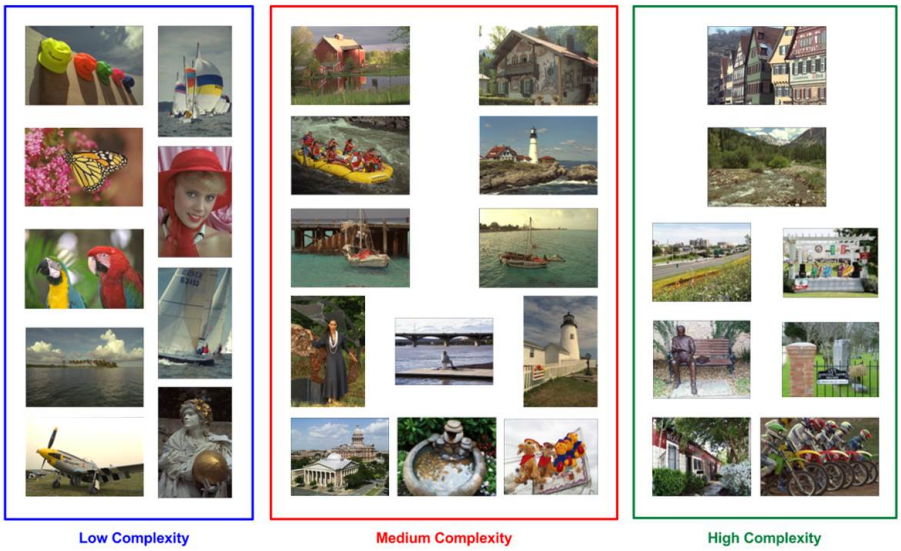

Figure 6.5: Original images from LIVE database grouped in three classes: high, medium and low complexity.

as low (medium) it changes to medium (high).

6.3.2 Applying the grouping strategy to better correlate data

Let us denote by y_i the subjective score (MOS/DMOS) for the i-th image of a given database, x_i the corresponding predicted score from an IQA metric and $f(x_i)$ the transformed metric value. For the transform function f logistic, polynomial or MR functions can be used.

Our proposal is to first classify the images within one of the following complexity groups: high (H), medium (M) or low (L) complexity and in a second step to perform the regression within each of these groups separately. Denoting by $\{y_i^L\}$, $\{y_i^M\}$ and $\{y_i^H\}$ the three sets of MOS and by $\{x_i^L\}$, $\{x_i^M\}$ and $\{x_i^H\}$ the corresponding predicted scores, the new proposal f_C for the transformation function is:

$$f_C(x) = \{f_L(x_i^L), f_M(x_i^M), f_H(x_i^H)\} \tag{6.4}$$

where f_L, f_M and f_H are the transformed functions found for each of the complexity groups separately.

As an example, we consider the NR metric by [94] and the JPEG-LIVE database. In figure 6.8 we plot each of the logisitic regression curves $f_L(x_i^L)$, $f_M(x_i^M)$ and $f_H(x_i^H)$. For a reference, also the function f is included in each of these subfigures. The three regression functions are also plot simultaneously. The corresponding PCC values obtained are as follows: $PCC(f) = 0.8893$, $PCC(f_L) = 0.9671$, $PCC(f_M) = 0.9476$, $PCC(f_H) = 0.9407$ and $PCC(f_C) = 0.9568$ respectively. With respect to $RMSE$ the values are: $RMSE(f) = 12.60$, $RMSE(f_L) = 7.77$, $RMSE(f_M) = 8.44$, $RMSE(f_H) = 7.72$ and $RMSE(f_C) = 8.01$ respectively. As expected, the strategy proposed has improved in terms of both PCC and $RMSE$ coefficients. Finally, to better visualize the comparison, the MOS vs the transformed metric using f and f_C are depicted in Figure 6.9.

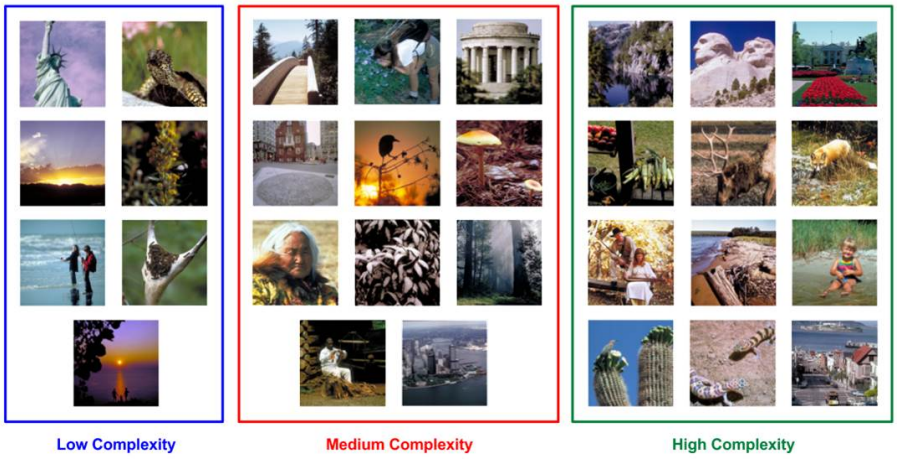

Figure 6.6: Original images from CSIQ database grouped in three classes: high, medium and low complexity.

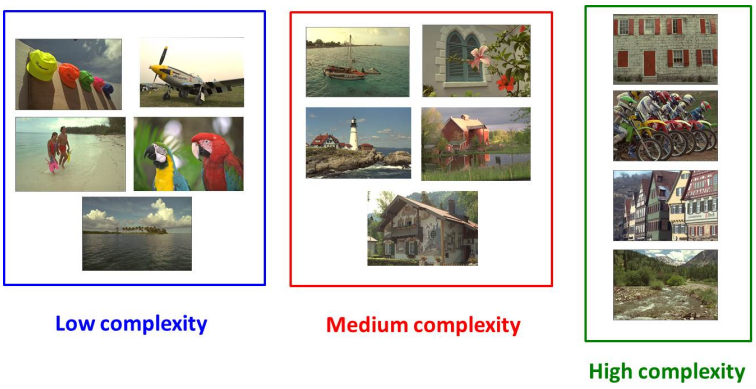

Figure 6.7: Original images from MICT database grouped in three classes: high, medium and low complexity.

Summarizing: images will be first classified according to their spatial complexity within one of the following groups: high, medium or low complexity. After this step, the correlation between objective and subjective data will be implemented within each group separately. Therefore, three regression functions will be obtained. For each of them, the correlation coefficients will be calculated and compared with the traditional case (no grouping). Experiments are performed on LIVE, MICT and CSIQ databases.

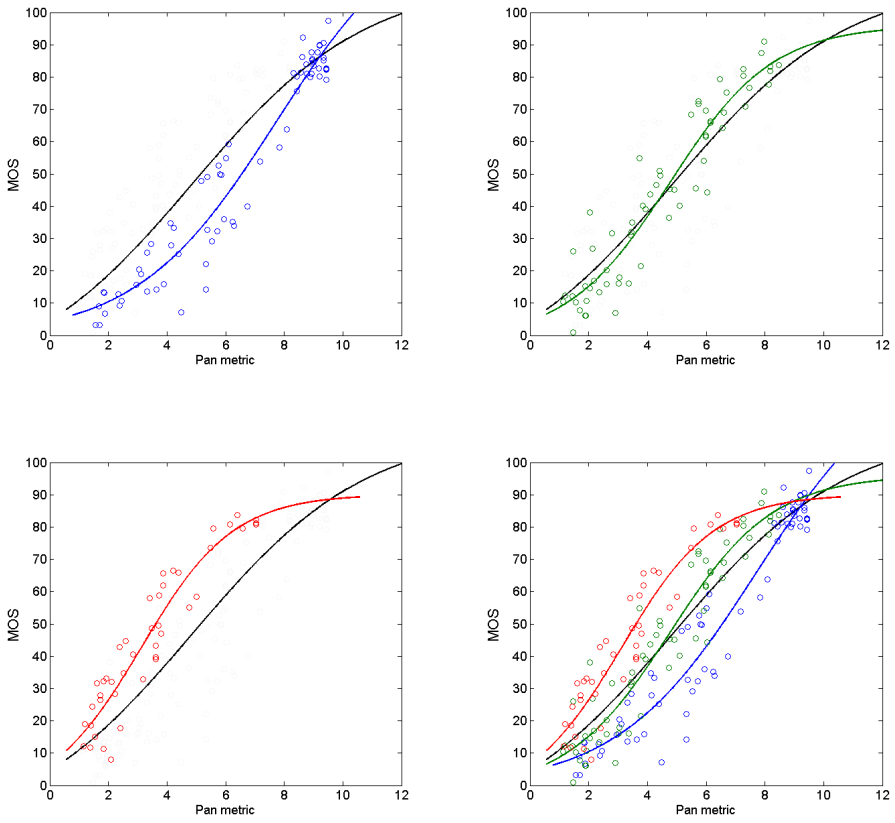

Figure 6.8: Logistic regression performed within each of the complexity groups: f_L (blue), f_M (green), f_H (red), f (black).

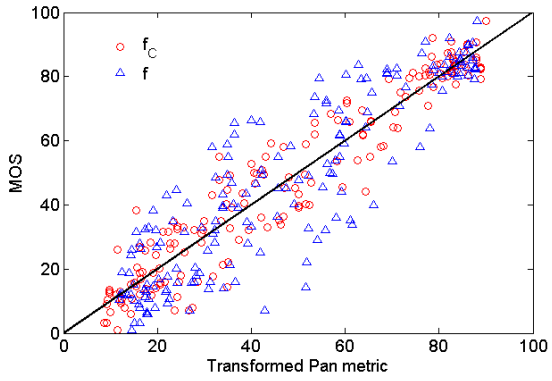

Figure 6.9: Transformed Pan metric: using f (triangles blue) and our proposal f_C (circles red)

6.3.3 Grouping strategy applied to NR metrics

Results on LIVE database for JPEG and white noise distortions

In Table 6.4 the PCC and $RMSE$ are reported for eleven JPEG-blockiness NR metrics for the case of JPEG - LIVE data. In the first and second row of the tables the values of the coefficients correspond to the use of Logistic Regression (LR) functions f and f_C respectively. In the third, fourth and fifth rows the coefficients are those corresponding to the transformed functions for each of the complexity groups, i.e. f_L, f_M and f_H respectively. Comparing each of the rows with respect to the first one, the best results for each metric are highlighted and colored.

	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11
PCC	PAN	VLA	WSB	WBE	GBIM	MUIJS	CHEN	BIQI	BRISQUE	BLIIND	NIQE
f	0,8893	0,9268	0,9787	0,9170	0,9703	0,9266	0,9455	0,9631	0,9866	0,9288	0,9526
f_C	**0,9568**	**0,9361**	**0,9811**	**0,9482**	**0,9792**	0,9194	**0,9775**	0,9607	**0,9881**	**0,9508**	**0,9667**
f_L	**0,9671**	**0,9706**	**0,9872**	**0,9605**	**0,9872**	0,8775	**0,9840**	0,9537	**0,9920**	**0,9386**	**0,9801**
f_M	**0,9476**	0,8898	0,9765	**0,9283**	**0,9713**	**0,9554**	**0,9671**	**0,9665**	0,9866	**0,9615**	0,9518
f_H	**0,9407**	0,9171	0,9706	**0,9460**	0,9691	**0,9570**	**0,9767**	**0,9658**	0,9784	**0,9598**	**0,9539**

	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11
RMSE	PAN	VLA	WSB	WBE	GBIM	MUIJS	CHEN	BIQI	BRISQUE	BLIIND	NIQE
f	12,60	10,35	5,66	10,99	6,66	10,37	8,98	7,41	4,49	10,22	8,38
f_C	**8,01**	**9,69**	**5,33**	**8,75**	**5,60**	10,84	**5,82**	7,65	**4,24**	**8,54**	**7,05**
f_L	**7,77**	**7,36**	**4,88**	**8,51**	**4,87**	14,67	**5,45**	9,20	**3,86**	10,55	**6,08**
f_M	**8,44**	12,05	5,69	**9,82**	**6,28**	**7,80**	**6,72**	**6,78**	**4,32**	**7,26**	**8,10**
f_H	**7,72**	**9,08**	**5,48**	**7,38**	**5,62**	**6,61**	**4,88**	**5,91**	4,70	**6,39**	**6,83**

Table 6.4: PCC and RMSE for JPEG LIVE Logisitic Regression

To better visualize and compare the results presented in Table 6.4 we depict in Figure 6.10 the bar plots of both PCC and RMSE corresponding to f and f_C. In the figure it is also shown the relative increase/decrease for PCC and the absolute difference for RMSE. That is, $\Delta PCC = [PCC(f_C) - PCC(f)]/PCC(f)$ and $\Delta RMSE = RMSE(f) - RMSE(f_C)$.

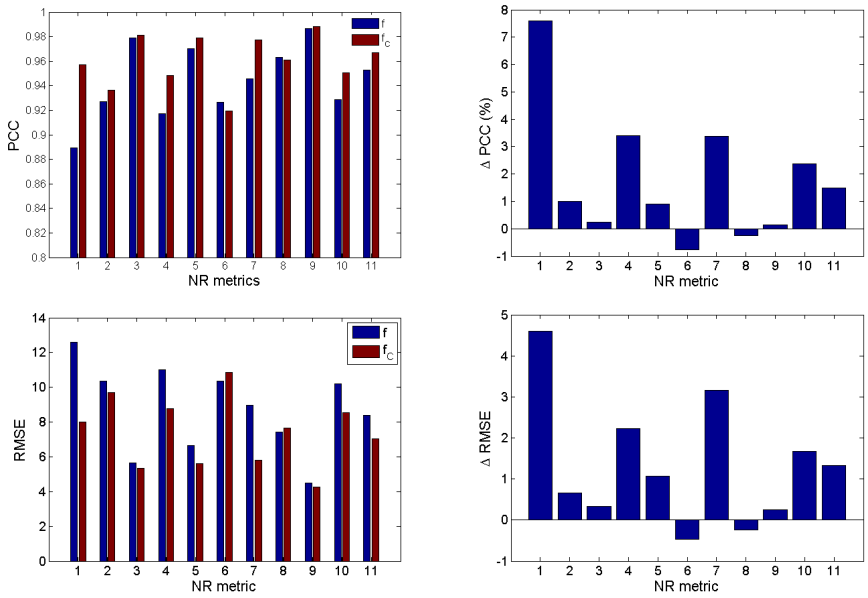

Figure 6.10: LIVE JPEG Logistic regression

From Table 6.4 and the corresponding Figure 6.10 we observe that applying the new transformation f_C, improvements are obtained in terms of both PCC and RMSE on nearly all the metrics, except for $M6$ and $M8$. The actual amount of performance gain, however, depends on the metric. The greatest improvement on PCC results equal to 7.5% for $M1$. Metrics $M4$, $M7$ and $M10$ show improvements around 3%. If we focus on the gain in performance when using f_H, f_M and f_L within the respectively complexity classes we note that metrics $M1$, $M4$, $M7$ and $M10$ show performance improvement overall three classes in terms of PCC. Similar conclusions are achieved if we analyze the RMSE coefficient.

In Table 6.5 and Figure 6.11 the results obtained using MR are shown. In this case we can observe that there is indeed a gain in performance in terms of both indexes for all metrics except for $M2$.

Comparing LR and MR transformations, we recall that the logistic function may vary if the initial parameters change. Therefore, the nonlinear optimization used in LR does not always lead to an exclusive result. On the other hand, as pointed out by Han et al. [44], the MR is not affected by any parameter choice and makes a unique minimum of the prediction error variance. The smaller the prediction error variance, the larger the PCC. Performance correlation indexes of MR would not be affected by any parameter. This fact could suggest MR as a fair benchmark for the performance comparison of the metrics.

	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11
PCC	PAN	VLA	WSB	WBE	GBIM	MUIJS	CHEN	BIQI	BRISQUE	BLIIND	NIQE
f	0,8966	0,9343	0,9811	0,9237	0,9732	0,9432	0,9492	0,9644	0,9866	0,9373	0,9545
f_C	**0,9609**	0,9343	**0,9842**	**0,9650**	**0,9818**	**0,9745**	**0,9769**	**0,9761**	**0,9899**	**0,9500**	**0,9677**
f_L	**0,9710**	**0,9717**	**0,9905**	**0,9605**	**0,9865**	**0,9828**	**0,9846**	**0,9850**	**0,9933**	0,9371	**0,9834**
f_M	**0,9509**	0,8832	**0,9824**	**0,9735**	**0,9768**	**0,9601**	**0,9618**	0,9641	**0,9898**	**0,9665**	0,9523
f_H	**0,9471**	0,9151	**0,9678**	**0,9576**	**0,9763**	**0,9756**	**0,9819**	**0,9715**	0,9791	**0,9492**	0,9487

	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11
RMSE	PAN	VLA	WSB	WBE	GBIM	MUIJS	CHEN	BIQI	BRISQUE	BLIIND	NIQE
f	12,21	9,82	5,33	10,56	6,33	9,16	8,67	7,29	4,49	9,60	8,22
f_C	**7,63**	9,83	**4,88**	**7,23**	**5,23**	**6,19**	**5,89**	**5,99**	**3,91**	**8,61**	**6,95**
f_L	**7,31**	**7,22**	**4,20**	**8,51**	**5,02**	**5,64**	**5,34**	**5,27**	**3,54**	10,68	**5,55**
f_M	**8,17**	12,39	**4,94**	**6,04**	**5,65**	**7,38**	**7,23**	**7,01**	**3,77**	**6,78**	**8,06**
f_H	**7,31**	**9,18**	5,73	**6,56**	**4,92**	**4,99**	**4,31**	**5,40**	4,63	**7,17**	**7,20**

Table 6.5: PCC and RMSE for JPEG LIVE Monotonic Regression

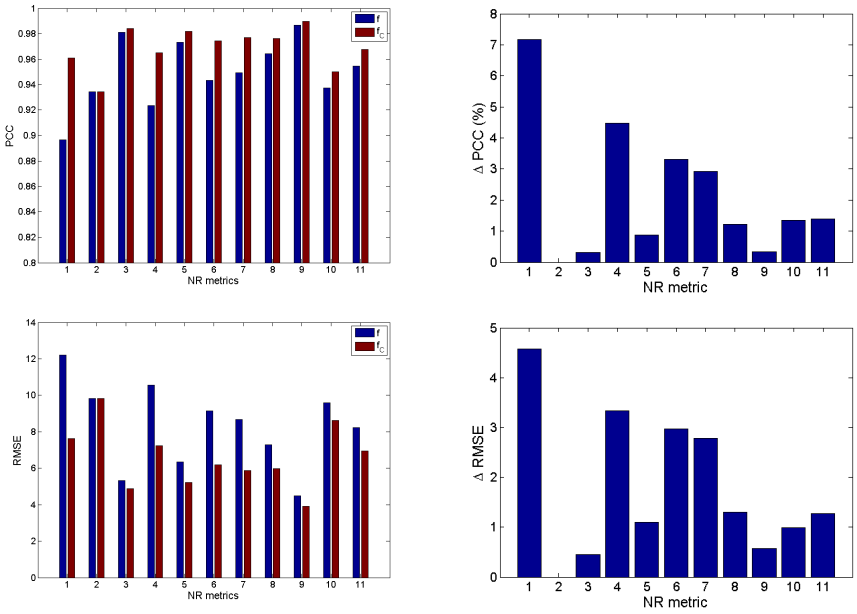

Figure 6.11: LIVE JPEG Monotonic regression

In Table 6.6 the grouping strategy is applied for the subset of the LIVE database distorted by white noise. In Figure 6.12 the comparisons are shown for PCC and RMSE indexes. The amount of performance gain in this case is small since the metrics themselves present high

values of PCC before applying the grouping strategy.

	M1	M2	M3	M4	M5
PCC	IMMERKAER	BIQI	BRISQUE	BLIIND	NIQE
f	0,9809	0,9930	0,9926	0,9652	0,9656
f_C	**0,9861**	**0,9935**	**0,9946**	**0,9732**	**0,9697**
f_L	0,8942	0,8277	0,9359	0,7032	0,8324
f_M	0,9330	**0,9734**	0,9529	0,8889	0,8256
f_H	0,9797	**0,9918**	**0,9936**	0,9603	0,9555

	M1	M2	M3	M4	M5
RMSE	IMMERKAER	BIQI	BRISQUE	BLIIND	NIQE
f	4,70	2,85	2,93	6,33	6,30
f_C	**4,02**	**2,76**	**2,51**	**5,56**	**5,92**
f_L	**2,97**	3,72	**2,34**	**4,72**	**3,68**
f_M	**3,95**	**2,52**	3,33	**5,03**	**6,19**
f_H	**4,15**	**2,65**	**2,33**	**5,77**	**6,11**

Table 6.6: PCC and RMSE for noise LIVE data and Logistic Regression

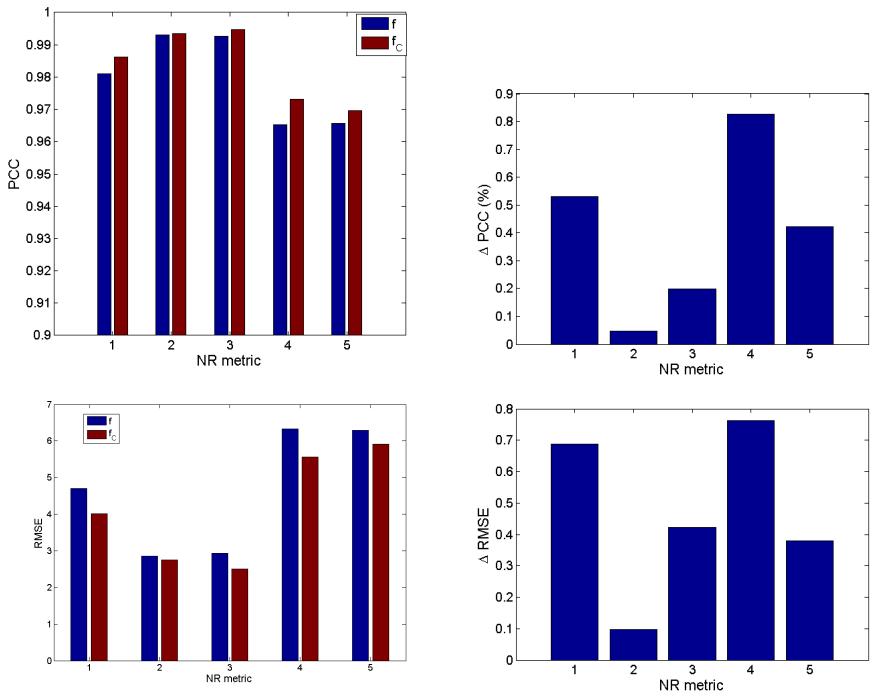

Figure 6.12: LIVE noise Logistic regression

Results on CSIQ and MICT database for JPEG

In Figure 6.13 we compare the results obtained for PCC and RMSE on the JPEG - CSIQ data and MR. The corresponding results for MICT database are reported in Figure 6.14.

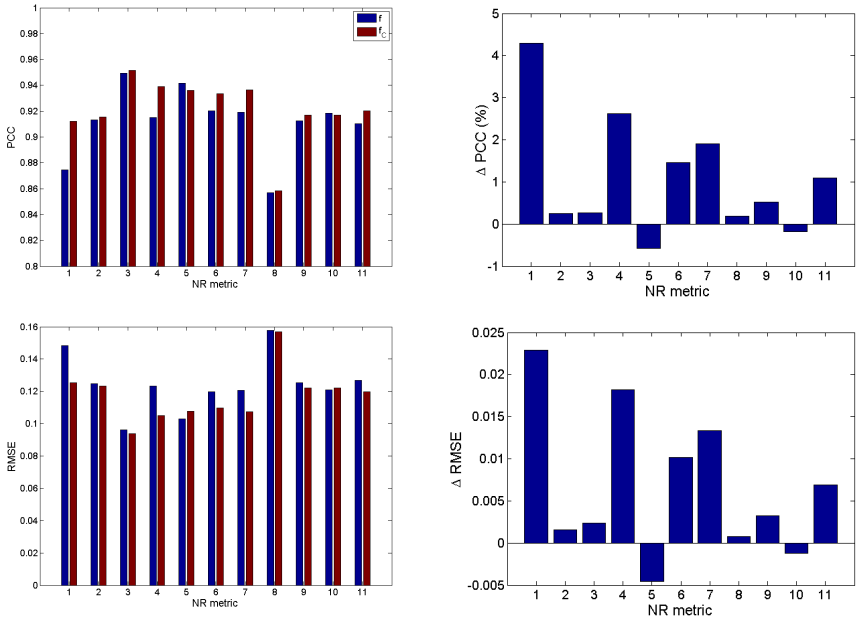

Figure 6.13: CSIQ JPEG Monotonic Regression

For CSIQ data, we observe increasements in PCC for nine out of the eleven metrics. The greatest perfomance improvement is again observed for metric $M1$, around 4%. The metrics $M4$, $M6$ $M7$ show smaller improvements, around 2%. For metrics $M5$ and $M10$ the performance is decreased.

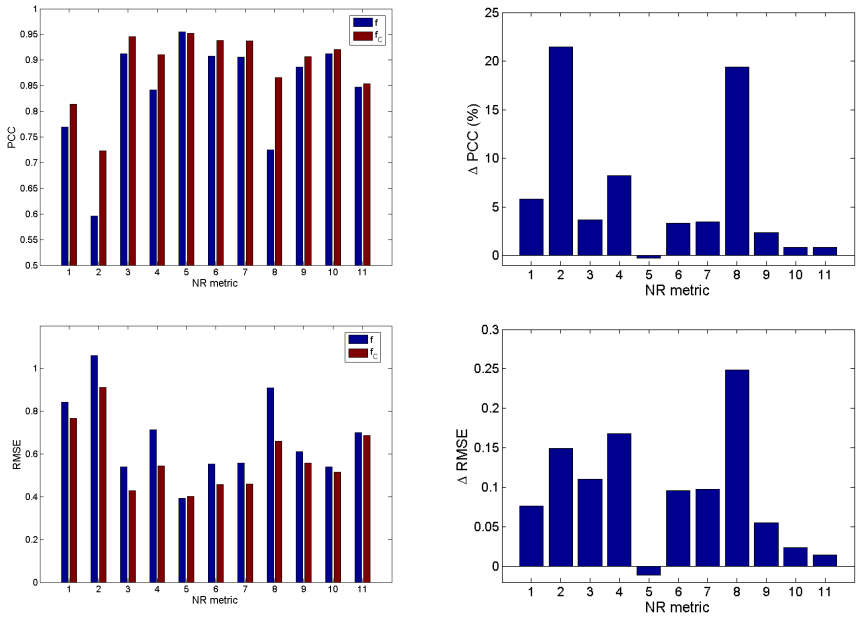

Figure 6.14: MICT JPEG Monotonic Regression

For the MICT data ΔPCC is always positive except for $M5$. An important gain is obtained for $M2$ and $M8$, around 20%.

Statistical significance

To evaluate the statistical significance of performance on both PCC and RMSE, we evaluate hypothesis testing. A variance-based hypothesis test is applied using the residuals between MOS and the quality predicted by the transformed metrics. The test is based on the assumption of Gaussianity of the residual differences. The F-statistic is used to compare the variance of two sets of sample points. The null hypothesis is that the residuals from one metric come from the same distribution and are statistically indistinguishable (with 95% confidence) from the residuals from another metric. A value of 0 indicates that applying f or f_C is statistically equivalent. Otherwise, the performance improvements are indicated as 1 and the performance decreases as "-1". The F-test scores are summarized for JPEG data in Table 6.7 for the three databases and two regression functions considered.

		M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11
DB	Reg.	PAN	VLA	WSB	WBE	GBIM	MUIJS	CHEN	BIQI	BRISQUE	BLIIND	NIQE
LIVE	LR	1	0	0	1	1	0	1	0	0	1	1
LIVE	MR	1	0	0	1	1	1	1	1	0	0	1
CSIQ	LR	0	1	1	0	0	0	1	0	0	1	0
CSIQ	MR	1	0	0	1	0	0	1	0	0	0	0
MICT	LR	0	0	1	0	0	0	1	0	0	0	0
MICT	MR	0	0	1	1	0	0	0	1	0	0	0

Table 6.7: F-test scores for NR metrics for JPEG data

It should, however, be noted that statistical significance testing is not straightforward, and the conclusions drawn from it largely depend e.g. on the number of sample points, on the selection of the confidence criterion, and on the assumption of normality of the residuals. These issues are extensively discussed in [116].

By referring to Table 6.7 we can see that introducing the grouping strategy gives always better or statistically equivalent results in terms of correlation coefficients. Analyzing the single NR metrics, the JPEG-blockiness specific $M7$ is the one that best benefited since improves for nearly all the datasets and regression functions (except MICT MR). On the other hand, for the general purpose $M9$ the results are always statistically equivalent. Comparing the databases, the grouping strategy seems to work better for LIVE data.

For the noise data, the F-test scores indicate that none of the improvements achieved is statistically significant.

6.3.4 Grouping strategy applied to FR metrics

In this section we perform again experiments on LIVE JPEG and white noise data but now focusing on FR methods. The MeTriX MuX Matlab package [1] was used for the evaluation of the following twelve FR metrics:

- Mean-Squared-Error (MSE), Signal Noise Ratio (SNR) and Peak Signal-to-Noise-Ratio (PSNR)
- Universal Quality Index (UQI) [138], Structural Similarity Index (SSIM) [140] and Multi-Scale SSIM index (MSSIM) [144]
- Visual Signal-to-Noise Ratio (VSNR) [19]
- Information Fidelity Criterion (IFC) [114], Visual Information Fidelity (VIF) and its pixel-domain version Pixel-based VIF (VIFP)
- Noise Quality Measure (NQM) [31] and Weighted Signal-to-Noise Ratio (WSNR).

We present results for PCC in Table 6.8 and Figure 6.15 for both JPEG and white noise data from LIVE.

	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12
PCC	MSE	SNR	PSNR	SSIM	MSSIM	UQI	VSNR	IFC	VIF	VIFP	NQM	WSNR
	JPEG											
f	0,8891	0,8768	0,8878	0,9502	0,9814	0,9072	0,9724	0,9444	0,9868	0,9814	0,9720	0,9658
f_C	**0,9551**	**0,9340**	**0,9775**	**0,9736**	**0,9831**	**0,9657**	**0,9769**	**0,9821**	0,9865	0,9781	**0,9764**	**0,9677**
f_L	**0,9838**	**0,9495**	**0,9830**	**0,9827**	**0,9904**	**0,9668**	**0,9779**	**0,9884**	**0,9953**	**0,9928**	**0,9866**	**0,9726**
f_M	**0,9390**	**0,9390**	**0,9790**	**0,9754**	**0,9821**	**0,9584**	**0,9781**	**0,9775**	0,9822	0,9784	0,9675	**0,9666**
f_H	**0,9366**	**0,8963**	**0,9651**	**0,9548**	0,9728	**0,9760**	0,9723	**0,9800**	0,9800	0,9529	**0,9754**	0,9609
	WHITE NOISE											
f	0,9826	0,9712	0,9858	0,9685	0,9849	0,9369	0,9779	0,9452	0,9663	0,9907	0,9869	0,9757
f_C	**0.9868**	**0,9726**	**0,9888**	**0,9805**	0,9790	**0,9670**	**0,9795**	**0,9772**	**0,9825**	0,9877	**0,9877**	**0,9763**
f_L	**0.9850**	**0,9802**	**0,9888**	**0,9896**	0,9819	**0,9391**	**0,9845**	**0,9682**	**0,9827**	**0,9914**	**0,9873**	0,9749
f_M	**0.9868**	0,9669	**0,9894**	**0,9727**	0,9761	**0,9785**	0,9720	**0,9811**	**0,9804**	0,9823	**0,9873**	0,9742
f_H	**0.9883**	0,9709	**0,9870**	**0,9806**	0,9787	**0,9797**	**0,9842**	**0,9804**	**0,9844**	**0,9912**	**0,9880**	**0,9800**

Table 6.8: PCC for FR methods, JPEG and white noise data from LIVE, Logistic Regression

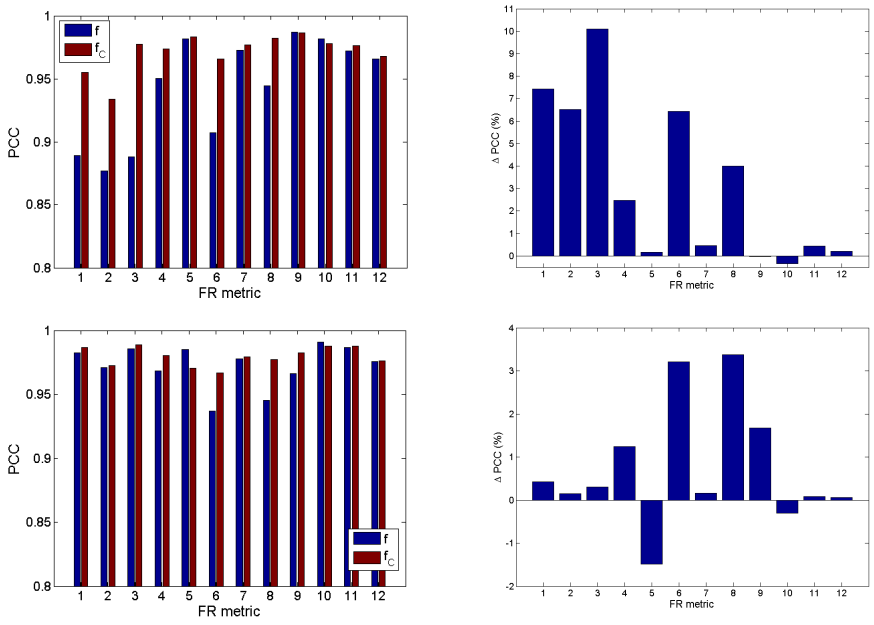

Figure 6.15: PCC for FR on LIVE data for JPEG and noise, logistic regression

Considering the results for JPEG distortion we observe that the greatest gain is obtained for the simplest and most widely used FR approaches like MSE, SNR and PSNR. Other more complete FR methods like SSIM and UQI that are based on image structure or natural scene statistics also show performance improvements. For the rest of the methods the

improvements are smaller, except for VIFP where a performance decrease is observed. The gains are lower in the case of noise distortion since the metrics themselves already show a high performance before applying the grouping strategy. On the other hand, MSSIM and VIFP show a decrease in PCC.

We conclude this section recalling that MSE, SNR and PSNR are appealing because are simple to calculate and are mathematically convenient in the context of optimization or real applications. However, they are widely criticized by the image quality community for their poor correlation with human perceived image quality. Applying the present complexity-based grouping strategy in the case of JPEG distortion can improve the simple methods' performances.

We have applied the complexity based strategy to all the distortions present in LIVE data and in Table 6.9 the statistical significant improvements obtained for the FR methods are reported.

M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	
MSE	SNR	PSNR	SSIM	MSSIM	UQI	VSNR	IFC	VIF	VIFP	NQM	WSNR	distortion
1	**1**	**1**	**1**	0	**1**	0	**1**	0	0	0	0	JPEG
1	0	0	**1**	**-1**	**1**	0	**1**	**1**	0	0	0	NOISE
1	**1**	**1**	0	0	**1**	0	**1**	0	0	0	0	JPEG2K
1	**1**	**1**	**1**	**-1**	0	0	0	**-1**	**1**	**1**	0	BLUR
0	**1**	**1**	0	**-1**	0	0	0	**1**	0	0	0	F-FADING

Table 6.9: Statistical significance tests for FR methods LIVE database - Logistic Regression

Analyzing Table 6.9 we confirm that the grouping strategy is more competitive with respect to the simpler metrics (PSNR and MSE) than for the more complex ones (MSSIM, VSNR, VIF, NQM, WSNR). The results depend on the distortion type: JPEG and blur are the ones that show better and statistically significant performances.

6.4 Conclusions

In order to better correlate objective and subjective data three different strategies have been proposed and tested on NR JPEG-blockiness metrics. The results are analyzed in terms of the correlation coefficients PCC and RMSE. The first proposal considers weighting the distorted image by its saliency map or its complement. Different computational models of saliency have been considered as well as the saliency map obtained from eye tracking experiments. However, no performance improvements are obtained. The second proposal weights each NR metric by a frequency dependent factor. Seven (five) NR JPEG-blockiness metrics our of nine show performance improvements in PCC when tested on LIVE (CSIQ) database. Finally, a grouping strategy based on the image spatial complexity is investigated. The proposal is tested on LIVE, MICT and CSIQ databases for JPEG and noise distortions. Also FR metrics are considered. The results depend on the NR metric considered and on the distortion. Statistical significance tests are performed. With respect to the JPEG we can say that there is indeed a gain in performance: improvements in PCC between 3 and 7% are observed for several metrics. For noise distortion the improvements are smaller since the

noise metrics themselves present a good performance before applying the grouping strategy. For the case of FR metrics, the greatest improvements (between 6 and 10 %) are obtained for the simplest FR metrics like MSE and PSNR.

Chapter 7

Conclusions

In this thesis we have discussed the current state of the art in IQA research. Among all the possible dimensions associated to define IQ, we have positioned here between the naturalness and fidelity dimensions. We have given a compendium of the available IQA methods, classifying and summarizing the different metrics. We have hypothesized an image workflow chain and outlined the relationship with the IQA, how and when the different kinds of metrics can be applied. The selection and use of the different metrics depend on the semantic content of the image, the application task, and the particularly imaging chain applied.

As possible application scenarios we had in mind quality monitoring. For example, evaluating the IQ of an image before printing it so that the final product reaches the desired quality level. For example, an IQA metric could be embedded within the printing workflow chain so that input images of high quality can be directly printed while low quality ones are discarded. Such application motivated by Océ, who founded this research work.

Different NR metrics are available as well as IQ databases for the validation process. Focusing on JPEG distortion, several metrics have been considered that show good performance in terms of correlation coefficients. However, there is a range of distortions where it is difficult for the metric to correctly predict the subjective scores. If the images are highly or slightly distorted, the metrics are in general able to describe the subjective scores but disagreements have been found in what we called "intermediate range" of distortions. Considering an image as a combination of content and distortion signals, in such intermediate range both content and distortion are significantly present and consequently, not easily decorrelated to be measured. It is therefore difficult for the metrics to measure the IQ with precision within the full range of possible distortions and also with respect to different image contents. The crosstalk between content and distortion signals influences both the subjective and objective quality assessment.

In general, the distortion ranges of the available databases vary from images of high quality to images highly corrupted. However, in real applications as the above mentioned it is not often to deal with so degraded images. Therefore, the IVL database was generated where the distortion range of JPEG and noise corrupted images has been chosen so as to represent real data as best as possible. Psychovisual experiments have been carried out with observers from our laboratory and Océ as well. The subjective scores have been collected for JPEG and noise distortions.

The principal contributions of this research work can be summarized as follows:

- The IQA field has been investigated focusing on a classification approach: we have

proposed to apply machine learning methods to classify images within three (classifier C3) and five (classifier C5) quality classes. The feature space was built with eleven NR metrics: seven specifically designed for JPEG-blockiness and four general purpose metrics. The assigned classes were obtained from the subjective scores. Different databases have been considered for both training and testing phases (LIVE, MICT and IVL). Training and testing C3 on IVL data, better performance was obtained compared to C5 when trained and tested on LIVE data. The better correlation of C3 with the subjective data could be attributed to the easier subjective task in the case of 3 classes. From the observer point of view it seems simpler to judge the quality of an image within one of 3 classes (Excellent, Fair or Bad) than making a finer analysis to classify it within 5 classes (Is the image excellent or good? Is the image poor or bad?). Observers are able to decide almost instantly whether a particular image is of good or poor quality but to quantify how good an image is, and the scale of quality to be used is more difficult.

- Different strategies to better correlate the objective and subjective data have been tested. Two of them, the frequency-based and complexity-based strategies, show performance improvements when considering NR metrics for JPEG-blockiness distortion. Different databases have been taken into account (LIVE, MICT, CSIQ). The complexity-based grouping proposal aims to correlate separately the data within three different groups of images, classified according to their spatial complexity in terms of low level image features. Even if the results depend on the metrics considered and the distortion type, there is indeed a gain in performance in terms of linear correlation and root mean square error.

The present research was concerned with single distortions. However, consumer images suffer in general of more than one distortion simultaneously. Although some IQA methods exist that address combining multiple distortions (most commonly noise and blur), few subjective studies have been performed on multiple distorted databases. Recently Jayaraman et al. [29] has presented a database of multiply distorted images, where two scenarios are considered: images first blurred and then JPEG compressed, and images first blurred and then corrupted by white Gaussian noise. Within this scenario, an IQA algorithm must not only consider the joint effects of the multiple distortions on the image, but also consider the effects of these distortions on each other.

As future research it will be considered the case of images corrupted by Gaussian noise and then JPEG compressed. The goal would be for example to evaluate if and how the quality perception of images corrupted by noise is modified in the presence of JPEG distortion. Testing the strategy proposals to other kind of distortions and NR metrics will be also addressed.

Bibliography

[1] *MeTriX MuX Visual Quality Assessment Package.* http://foulard.ece.cornell.edu/gaubatz/metrixmux/.

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

[3] http://www.oce.com, 2013.

[4] A. K. Moorthy A. Mittal and A. C. Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on ImageProcessing*, 21:4695–4708, 2012.

[5] P. L. Callet A. Ninassi, O. L.Meur and D. Barba. which semilocal visual masking model forwavelet based image quality metric?. In *Proceedings of the 15th IEEE International Conference on Image Processing*, pages 1180–1183, 2008.

[6] R. Achanta, F. Estrada, P. Wils, and S. Susstrunk. Salient region detection and segmentation. In *Computer Vision Systems, ser. Lecture Notes in Computer Science 5008*, pages 66–75. Springer Berlin / Heidelberg, 2008.

[7] E. Allen, S. Triantaphillidou, and R. E. Jacobson. Image quality comparison between JPEG and JPEG2000. i. psychophysical investigation. *The Journal of imaging science and technology*, 51:548–258, 2007.

[8] C. Bartleson. The combined influence of sharpness and graininess on the quality of color prints. *Journal Photogr Sci*, 30:33–38, 1982.

[9] P. Bartlett, S. Boucheron, and G. Lugosi. Model selection and error estimation. *Machine Learning*, pages 85–113, 2002.

[10] BCR's CDP Digital Imaging Best Practices Working Group. BCR's CDP Digital Imaging Best Practices Version 2.0. Technical report, 2008.

[11] A.C. Bovik and S. 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.

[12] T. Brandao and M. Queluz. No-reference image quality assessment based on dct domain statistics. *Signal Processing*, 88(4):822 – 833, 2008.

[13] L. Breiman, J. Friedman, and C. Olshen, R. Stone. *Classification and Regression Trees*. Wadsworth and Brooks/Cole, 1984.

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

[15] 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.

[16] M. Chacon-Murguia, A. Corral-Saenz, and R. Sandoval-Rodriguez. Image complexity measure: A human criterion free approach. In *Proc. Annual Meeting of the North American Fuzzy Information Processing Society (NAFIPS)*, pages 241–246, 2005.

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

[18] D. M. Chandler. Seven challenges in image quality assessment: Past, present, and future research. *ISRN Signal Processing*, 23:Article ID 905685, 53 pages, 2013.

[19] D.M. Chandler and S.S. Hemami. Vsnr: A wavelet-based visual signal-to-noise ratio for natural images. *IEEE Transactions on Image Processing*, 16:2284–2298, 2007.

[20] C. Charrier, O. Lezoray, and G. Lebrun. Machine learning to design full-reference image quality assessment algorithm. *Signal Processing: Image Communication*, 27:209–219, 2012.

[21] C. Chen and A. Bloom. A blind reference-free blockiness measure. In *Lecture Notes in Computer Science*, volume 6297, pages 112–123. Springer-Verlag Berlin Heidelberg, 2010.

[22] 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.

[23] 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.

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

[25] S. Corchs, F. Gasparini, F. Marini, and R. Schettini. Image quality: a tool for no-reference assessment methods. In *Image Quality and System Performance VIII, IS&T/SPIE Electronic Imaging*, volume 7867, pages 78760X (1–9). SPIE, 2011.

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

[27] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20:273–297, 1995.

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

[29] A. Moorthy D. Jayaraman, A. Mittal and A. Bovik. Objective quality assessment of multiply distorted images. In *Proc. of the Asilomar Conference on Signals, Systems and Computers*, 2012.

[30] 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.

[31] N. Damera-Venkata, T. Kite, W. Geisler, B. Evans, and A. Bovik. Image quality assessment based on a degradation model. *IEEE Transactions on Image Processing*, 9:636–650, 2000.

[32] R. Datta, D. Joshi, J. Li, and J. Wang. Studying aesthetics in photographic images using a computational approach. In *In Proc. ECCV*, pages 7–13, 2006.

[33] 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.

[34] EC. *Europeans guidelines on quality criteria for computed tomography*, (accessed Frebruary 09, 2012).

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

[36] F. Frey and J. Reilly. Digital imaging for photographic collections: foundations for technical standards. Technical report, 1999.

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

[38] S. Gabarda and G. Cristobal. No-reference image quality assessment through the von mises distribution. *J. Opt. Soc. Am. A*, 29(10):2058–2066, 2012.

[39] F. Gasparini, M. Guarnera, F. Marini, and R. Schettini. No-reference metrics for demosaicing. In *Proc. of the SPIE*, pages vol 7529, p 752911, 2010.

[40] P. Gastaldo and J. Redi. Machine learning solutions for objective visual quality assessment. In *Sixth international workshop on Video Processing and Quality Metrics (VPQM) 2012*, 2012.

[41] B. Girod. Digital images and human vision. chapter What's wrong with mean-squared error?, pages 207–220. MIT Press, Cambridge, MA, USA, 1993.

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

[43] N. Joshi H. Tang and A. Kapoor. Learning a blind measure of perceptual image quality. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 305–312, 2011.

[44] Y. Han, Y. Cai, Y. Cao, and X. Xu. Monotonic regression: A new way for correlating subjective and objective ratings in image quality research. *IEEE Transactions on Image Processing*, 21:2309–2313, 2012.

[45] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. In *Advances in Neural Information Processing Systems 19*, pages 545–552. MIT Press, 2007.

[46] D. Hasler and S. Süsstrunk. Measuring colorfulness in natural images. volume 5007, pages 87–95. SPIE, 2003.

[47] X. Hou and L. Zhang. Saliency detection: A spectral residual approach. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007.

[48] C. Hsu, C. Chang, and C. Lin. A practical guide to support vector classification. *Bioinformatics*, 1:1–16, 2010.

[49] I3A. *Fundamentals and review of considered test methods.* CPIQ Initiative Phase 1 White Paper, 2007.

[50] Imatest. *Digital Image Quality Testing.* http://www.imatest.com, 2010.

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

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

[53] ISO. *Image technology colour management - Architecture, profile format and data structure ? Part 1: Based on ICC.1:2004-10. ISO 15076-1.* 2005.

[54] ISO. *ISO 12233 Chart Data*, (accessed Frebruary 09, 2012).

[55] L. Itti and C. Koch. Computational modelling of visual attention. *Nature Reviews Neuroscienc*, 2:194–203, 2001.

[56] ITU. Methodology for the subjective assessment of the quality for television pictures. Technical report, ITU-R Rec. BT. 500-13 (01.12), 2012.

[57] P. Gastaldo R. Zunino J. Redi, H. Liu and I. Heynderickx. How to apply spatial saliency into objective metrics for jpeg compressed images? In *Proc. of the IEEE Int. Conf. Image Processing*, pages 961–964, 2009.

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

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

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

[61] B. W. Keelan. *Handbook of Image Quality: Characterization and Prediction.* 2002.

[62] T.M. Kusuma and H.-J. Zepernick. A reduced-reference perceptual quality metric for in-service image quality assessment. In *Mobile Future and Symposium on Trends in Communications, 2003. SympoTIC '03. Joint First Workshop on*, pages 71 – 74, 2003.

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

[64] E. Larson and D. Chandler. Most apparrent distortion: full reference image quality assessmente and the role of strategy. *Journal of Electronic Imaging*, 19:011006 1–21, 2010.

[65] 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.

[66] C. Li, A. Bovik, and X. Wu. Blind image quality assessment using a general regression neural network. *IEEE Transactions on Neural Networks*, 22:793799, 2011.

[67] Q. Li and Z. 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.

[68] W. Lin and C. Jay Kuo. Perceptual visual quality metrics: A survey. *J. Visual Communication and Image Representation*, 22:297–312, 2011.

[69] H. Liu and I. Heynderickx. A perceptually relevant no-reference blockiness metric based on local image characteristics. *EURASIP Journal on Advances in Signal Processing*, pages ID 263540,1–14, 2009.

[70] H. Liu and I. Heynderickx. Studying the added value of visual attention in objective image quality metrics based on eye movement data. In *Proc. of the IEEE International Conference on Image Processing*, pages 3097–3100, 2009.

[71] T. Liu, W. Lin, and C. Jay Kuo. Image quality assessment using multi-method fusion. *IEEE Trans. Image Processing*, 22:1793–1807, 2013.

[72] 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.

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

[74] Q. Ma, L. Zhang, and B. Wang. New strategy for image and video quality assessment. *J. Electronic Imaging*, 19(1):011019, 2010.

[75] 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.

[76] F. Marini, C. Cusano, and R. Schettini. No-reference metrics for jpeg: analysis and refinement using wavelets. In *Proc. SPIE 2010*, volume 7529, pages 75290C–1–75290C–9, 2010.

[77] P. Marziliano, F. Dufaux, S. Winkler, and T. Ebrahimi. A no-reference perceptual blur metric. In *IEEE 2002 International Conference on Image Processing*, pages 57–60, 2002.

[78] A. Mittal, R. Soundararajan, and A. C. Bovik. Making a completely blind image quality analyzer. *IEEE Signal Processing Letters*, 20:209–212, 2013.

[79] M. Miyahara, K. Kotani, and V. Algazi. Objective picture quality scale (pqs) for image coding. *IEEE Trans. Commun*, 46:12151225, 1998.

[80] A. Moorthy and A. Bovik. Visual quality assessment algorithms: what does the future hold? *Multimedia Tools and Applications*, 51:675–696, 2011.

[81] A. Moorthy and A. C. Bovik. A two-step framework for constructing blind image quality indices. *IEEE Signal Processing Letters*, 17:513–516, 2010.

[82] 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.

[83] R. Muijs and I. Kirenko. A no-reference blocking artifact measure for adaptive video processing. In *Proceedings of the 13th European Signal Processing Conference 2005*, 2005.

[84] M. Narwaria and W. Lin. Objective image quality assessment based on support vector regression. *IEEE Trans. Neural Networks*, 21:515–519, 2010.

[85] M. Narwaria, W. Lin, and A.E. Cetin. Scalable image quality assessment with 2d mel-cepstrum and machine learning approach. *Pattern Recognition*, 45:299–313, 2012.

[86] 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.

[87] M. Nishiyama, T. Okabe, I. Sato, and Y. Sato. Aesthetic quality classification of photographs based on color harmony. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 33–40, june 2011.

[88] K. H. Oh, S. Triantaphillidou, and R.E. Jacobson. Scene classification with respect to image quality measurements. In *Proc. SPIE*, volume 7529, pages 752908–752908–10, 2010.

[89] A. Oliva, M. Mack, M. Shrestha, and A. Peeper. Identifying the perceptual dimensions of visual complexity of scenes. In *Proc. 26th Annual Meeting of the Cognitive Science Society*, 2004.

[90] S. Olwen. Noise variance estimation in images. In *Proc. of the 8th SCIA, Tromso, Norway*, 1993.

[91] 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.

[92] E. Ong, X. Yang, W. Lin, Z. Lu, S. Yao, X. Lin, S. Rahardja, and C. Boon. Perceptual quality and objective quality measurements of compressed videos. *J. Vis. Commun. Image Representation*, 17:717–737, 2006.

[93] N. Otsu. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9:62–66, 1979.

[94] 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. *Signal Processing: Image Communication*, 19(6):499 – 506, 2004.

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

[96] J. Perkio and A. Hyvarinen. Modelling image complexity by independent component analysis, with application to content-based image retrieval. In *Proc. 19th International Conference on Artificial Neural Networks (ICANN)*, volume 2, pages 704–714, 2009.

[97] 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.

[98] F. Questier, R. Put, D. Coomans, B. Walczak, and Y. Vander Heyden. The use of cart and multivariate regression trees for supervised and unsupervised feature selection. *Chemometrics and Intelligent Laboratory Systems*, 76:45–54, 2005.

[99] 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.

[100] J. Redi, P. Gastaldo, I. Heynderickx, and R. Zunino. Color distribution information for the reduced-reference assessment of perceived image quality. *IEEE Trans. CSVT*, 20:1757–1769, 2010.

[101] J. Riagau, M. Feixas, and M. Sbert. An information-theoretic framework for image complexity. In *Proc. Computational Aesthetics in Graphics, Visualization and Imaging*, pages 177–184, 2005.

[102] R. Subhabrata S. Bhattacharya and M. Shah. A holistic approach to aesthetic enhancement of photographs. *ACM Trans. Multimedia Comput. Commun. Appl.*, 7S:21:1–21:21, November 2011.

[103] H. V. Zhao S. Tjoa, W. S. Lin and K. J. R. Liu. Block size forensic analysis in digital images. In *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 633–636, 2007.

[104] M. Saad, A. Bovik, and C. Charrier. Blind image quality assessment: A natural scene statistics approach in the dct domain. *IEEE Transactions on Image Processing*, 21:3339–3352, 2012.

[105] 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.

[106] 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.

[107] 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.

[108] R. Schettini, C. Brambilla, G. Ciocca, A. Valsasna, and M. De Ponti. A hierarchical classification strategy for digital documents. *Pattern Recognition*, 35:1759–1769, 2002.

[109] R. Schettini, C. Brambilla, C. Cusano, and G. Ciocca. Automatic classification of digital photographs based on decision forest. *International Journal of Pattern Recognition and Artificial Intelligence*, 18:819–845, 2004.

[110] R Schettini and F Gasparini. A review of redeye detection and removal in digital images through patents. *Recent Patents on Electrical Engineering*, 2(1):45 – 53, 2009.

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

[112] G. Sharma. *Digital Color Imaging Handbook*, volume 29. CRC, 2003.

[113] H. Sheikh and A. Bovik. Image information and visual quality. *IEEE Transactions on Image Processing*, 15:430–444, 2006.

[114] H. Sheikh, A. Bovik, and G. de Veciana. An information fidelity criterion for image quality assessment using natural scene statistics. *IEEE Transactions on Image Processing*, 14:2117–2128, 2005.

[115] 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.

[116] 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*, 15:3440–3451, 2006.

[117] 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.

[118] E. Simoncelli and B. Olshausen. Natural image statistics and neural representation. *Annual Review of Neuroscience*, 24(1):1193–1216, 2001.

[119] R. Soundararajan and A.C. Bovik. Rred indices: Reduced reference entropic differencing for image quality assessment. *IEEE Transactions on Image Processing*, 21(2):517–526, 2012.

[120] S. Suresh, V. Babu, and H. Kim. No-reference image quality assessment using modified extreme learning machine classifier. *Applied Soft Computing*, 9:541–552, 2009.

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

[122] H. Tang, N. Joshi, and A. Kapoor. Learning a blind measure of perceptual image quality. In *Proc. of the International Conference on Computer Vision and Pattern Recognition*, 2011.

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

[124] P. Teo and D. Heeger. Perceptual image distortion. In *in Proc. SPIE*, pages 982–986, 1994.

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

[126] H. Tong, M. Li, C. Zhang, H.and Zhang, J. He, and W. Ma. Learning no-reference quality metric by examples. In *Proc. Multimedia Modelling Conference 2005*, pages 247 – 254, 2005.

[127] H. Tong, M. Li, H. Zhang, and C. Zhang. Learning no reference quality metric by examples. In *Proc. of the 11th International MultiMedia Modelling Conference*, 2005.

[128] Y. Tong, H. Konik, F. Cheikh, and A. Tremeau. Full reference image quality assessment based on saliency map analysis. *Journal of Imaging Science*, 54(3):30503–1–30503–14, 2010.

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

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

[131] S. Tourancheau, F. Autrusseau, Z.M.P. Sazzad, and Y. Horita. Impact of subjective dataset on the performance of image quality metrics. In *Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on*, pages 365–368, 2008.

[132] S. Triantaphillidou, E. Allen, and R. E Jacobson. Image quality of jpeg vs jpeg 2000, part 2: Scene dependency, scene analysis and classification. 51:259–270, 2007.

[133] H.-J. Zepernick U. Engelke, H. Kaprykowsky and P. Ndjiki-Nya. Visual attention in quality assessment. *IEEE Signal Processing Magazine*, 28:50–59, 2011.

[134] V. Vapnik. *The Nature of Statistical Learning Theory.* Springer Verlag, 1995.

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

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

[137] M. Wainwright, O. Schwartz, and E. Simoncelli. Natural image statistics and divisive normalization: Modeling nonlinearity and adaptation in cortical neurons. In *Statistical Theories of the Brain, MIT Press*, 2001.

[138] Z. Wang and A. Bovik. A universal image quality index. *IEEE Signal Processing Letters*, 9:81–84, 2002.

[139] Z. Wang and A. Bovik. Mean squared error: Love it or leave it? a new look at signal fidelity measures. *Signal Processing Magazine, IEEE*, 26(1):98 –117, 2009.

[140] 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.

[141] Z. Wang, A.C. Bovik, and B.L. Evans. Blind measurement of blocking artifacts in images. In *in Proc. IEEE Int. Conf. Image Proc.*, pages 981–984, 2000.

[142] Z. 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 477–480. IEEE, 2002.

[143] Z. Wang and E. 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.

[144] Z. Wang, E. Simoncelli, and A. Bovik. Multi-scale structural similarity for image quality assessment. In *37th IEEE Asilomar Conference on Signals, Systems and Computers*, 2003.

[145] Z. Wang, E. Simoncelli, and H. Hughes. Local phase coherence and the perception of blur. In *in Adv. Neural Information Processing Systems (NIPS03*, pages 786–792. MIT Press, 2004.

[146] A. Watson, R. Borthwick, and M. Taylor. Image quality and entropy masking. In *SPIE Human Vision and Electronic Imaging Conference*, volume 3016, pages 2–12, 1997.

[147] 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.

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

[149] C. Wee, R. Paramesran, R. Mukundan, and X. Jiang. Image quality assessment by discrete orthogonal moments. *Pattern Recognition*, 43(12):4055 – 4068, 2010.

[150] S. Winkler. Analysis of public image and video databases for quality assessment. *IEEE J. Selected Topics in Signal Processing*, 6:616–625, 2012.

[151] S. Winkler and S. 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.

[152] H. Wu and M. Yuen. A generalized block-edge impairment metric for video coding. *IEEE Signal Processing Letters*, 4:317–320, 1997.

[153] X-Rite. *X-Rite ColorChecker Classic*, (accessed Frebruary 09, 2012).

[154] P. Ye and D. Doermann. No-reference image quality assessment using visual codebooks. *IEEE Transactions on Image Processing*, 21(7):3129–3138, 2012.

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

[156] J. You, A. Perkis, M. Hannuksela, and M. 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.

[157] H. Yu and S. Winkler. Image complexity and spatial information. In *Proc. 5th International Workshop on Quality of Multimedia Experience (QoMEX)*, 2013.

[158] Z. Yu, H. Wu, S. Winkler, and T. Chen. Vision-model-based impairment metric to evaluate blocking artifacts in digital video. In *Proc. of the IEEE 90*, pages 154–169, 2002.

[159] 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.