Contents

List of Figures

List of Tables

Chapter 1

Objective Quality Assessment Metrics

Contents

1.1 Introduction

In the past years, the development of novel image and video coding technologies has spurred the interest in developing digital video communications. The definition of evaluation mechanisms to assess the video quality plays a major role in the overall design of video communication systems.

As [2] explains, the image quality measurement is very important for most image processing applications. An image quality metric has mainly three kinds of applications:

It can be used to monitor image quality as for example in an image and video acquisition system which can use the quality metric to monitor and automatically adjust the system to obtain the best quality. Or also a network video server can use it to examine the quality of the digital video transmitted and control the video streaming. It can be also employed to benchmark image processing systems, algorithms and encoder proposals. And it can be embedded into an image processing system to optimize the algorithms and the parameter settings. For instance, in a visual communication system, a quality metric can help optimal design of the prefiltering and bit assignment algorithms at the encoder and the postprocessing algorithms at the decoder.

The most reliable way of assessing the quality of a video or image is subjective evaluation, because human beings are the ultimate receivers in most applications. But this way of assess image quality is not appropriate for the mentioned applications.

The Mean Opinion Score (MOS), which is a subjective quality metric obtained from a number of human observers, has been regarded for many years as the most reliable form of quality measurement. However in order to achieve statistically relevant results, the MOS method has to evaluate a huge test population, so it is too cumbersome, time consuming, not suited for real-time and is expensive for most applications.

The MOS, is generated by averaging the results of a set of subjective tests, where a number of viewers rate the image or video quality of the presented images or sequences, by the way of one of the standardized methodologies proposed in the following international recommendations:

- ITU-R BT.500-11 (2002) & ITU-R BT.500-12 (09/2009) [3, 4] Methodology for the subjective assessment of the quality of television pictures: This Recommendation provides methodologies for the assessment of picture quality including general methods of test, the grading scales and the viewing

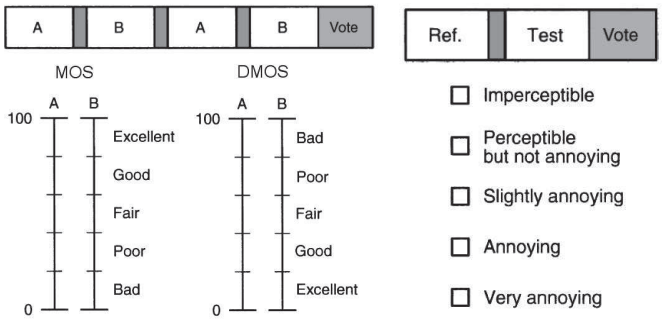


Figure 1.1: Presentation sequence and rating scale for (a) DSCQS (b) DSIS, methods

conditions. It recommends the Double-Stimulus Impairment Scale (DSIS) method and the Double-Stimulus Continuous Quality-Scale (DSCQS) method as well as alternative assessment methods such as Single-Stimulus (SS) methods, stimulus-comparison methods, Single Stimulus Continuous Quality Evaluation (SSCQE) and Simultaneous Double Stimulus for Continuous Evaluation (SDSCE) method.

- ITU-T P.910 (04/2008) [5] Subjective video quality assessment methods for multimedia applications: Describes non-interactive subjective assessment methods for evaluating the one-way overall video quality for multimedia applications such as videoconferencing, storage and retrieval applications, tele-medical applications, etc.

The three classes of subjective assessment methologies: single stimulus methods, comparison methods and double stimulus methods, recommended in these standards are briefly summarized below.

- Double Stimulus Continuous Quality Scale (DSCQS): The reference and the distorted image (or sequence) are presented twice to the viewer alternating between reference and distorted versions, see Figure 1.1(a). The viewers should rank the perceived quality in a continuous scale of 0-100 (being 0 bad and 100 excellent). Multiple pairs of reference and distorted images (or sequences) are shown to the viewers but they are not told which one is the reference or the distorted one. Analysis is based on the difference in rating for each pair, which is often calculated from an equivalent numerical scale from 0 to 100. In the case of DSCQS, the Difference Mean Opinion Score (DMOS) could be used instead of MOS. It consists of the mean of differential subjective

scores. For each viewer and image (or sequence) the raw scores are first converted to difference scores, that is, the difference between the given score to the reference and distorted version. These scores are further normalized as explained in [6] to obtain Zscores [7] that are finally rescaled to the 0-100 range to obtain the DMOS score for that image or sequence, where 0 represents the best quality value (no difference between reference and distorted image) .

- Double Stimulus Impairment Scale (DSIS). Unlike DSCQS, the viewers know which one is the reference image (or sequence), that is presented first, followed by the distorted one. In DSIS variant II this presentation is repeated once. The viewers rate the images/sequences in the five-level scale presented in Figure 1.1. This method is named as Degradation Category Rating (DCR) in the ITU-T P.910.

- Single Stimulus Continuous Quality Evaluation (SSCQE). Here the viewers are only shown the distorted image/sequence, but for a longer duration than in the previous methods, typically 20-30 minutes, and rate simultaneously while watching the sequence the perceived quality using a slider in the same scale that DSCQS.

- Absolute Category Rating (ACR). Like SSCQE is a single stimulus with only the distorted version showed to the viewers. They provide a single quality rate for the overall sequence using the five-level scale from Fig. 1.1(a).

- Pair Comparison (PC). This method pairs the references and distorted versions in any possible combination of compression degree and or used encoder. The pair is shown twice in rapid succession and at the end the viewer should choose which version of the pair has better quality.

These methods generally have different applications. DSCQS is is the preferred method when the quality of test and reference sequence are similar, because it is quite sensitive to small differences in quality. The DSIS method is better suited for evaluating clearly visible impairments such as artifacts caused by transmission errors, for example. As for all subjective tasks, different results can be achieved depending on how the video or image content is presented to the viewers and which method is used. Inclusive for the same content the way and the order in which it is presented to the viewers can bias the results in a desired direction.

In [8, 9, 10] authors review and compare some of these standardized testing methodologies, emphasizing the benefits and problems of each method. They analyzed the results of SSCQE and DSCQS methods, concluding that high

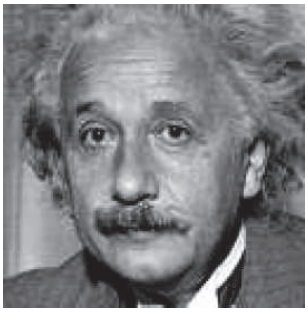
correlated results between these methods can be achieved if the SSCQE duration of the sequences is reduced to 9 to 15 seconds. Their experiments conclude that the participating viewers considered at most the last 9 to 15 seconds of video when forming their quality estimate. This is not to say that long sequences are completely without other merits. Nonetheless, when long video sequences are used in SSCQE tests, test designers should not necessarily expect a panel of viewers to rate the video inherently differently than if shorter sequences are used. The advantages of using SSCQE as a substitute of DSCQS for video comparisons, would include faster testing (or more clips rated for the same amount of viewing time spent) and less viewer fatigue.

Another comparison of the DSCQS and DSIS II scales can be found in [11, 10] where authors study the effects of context in the different methods. One type of contextual effects is created when there are fluctuations in the subjective rating of sequences based on the types and amount of impairments presented in the preceding sequence in the test. For example, a sequence with moderate impairment that follows a set of sequences with weak impairment may be judged lower in quality than if it follows sequences with strong impairment. A common method used to try and counterbalance this type of contextual effect is the randomization of the test trial presentation order. Using it, they finally conclude that the DSCQS method has reduced contextual effects, being the best method to use in order to minimize contextual effects for subjective picture quality assessment.

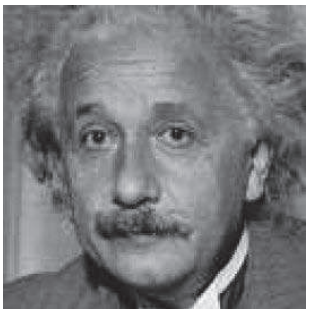
These aforementioned studies reveal that the selection of the proper method for presenting the references and the distorted versions of our images or sequences could result in varying results. Besides, we have to take into account the time needed to prepare the test images, the distorted versions, the ordering of the test sequences, the viewing conditions and to be able to enroll sufficient viewers to have statistically representative results.

Traditionally, in order to avoid the need to perform such time consuming subjective tests, the scientific community mostly has used the Mean Square Error (MSE) and the Peak to Noise Ratio (PSNR) to assess quality and compare the performance of different and competing encoding proposals. This is because MSE and consequently PSNR has many attractive features [12], it is simple to calculate and parameter free, mathematically easy to deal for optimization purposes, is the natural way to define the energy of the error signal and finally is the most commonly used metric. Technically, MSE measures image difference, whereas PSNR measures image fidelity, i.e. how closely an image resembles a reference image, usually the uncorrupted original. Due to the popularity of these metrics, most of the results from previous comparison works are expressed with

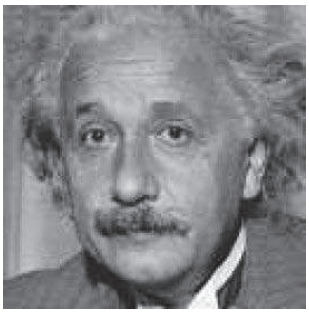
them, because using it saves time and effort while comparing, and as a side effect, it further propagates the use of MSE and PSNR.



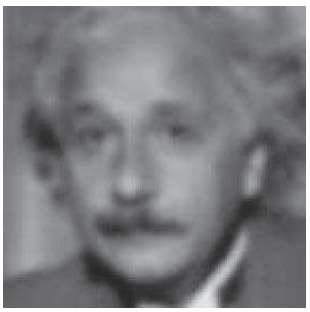
(a) Original



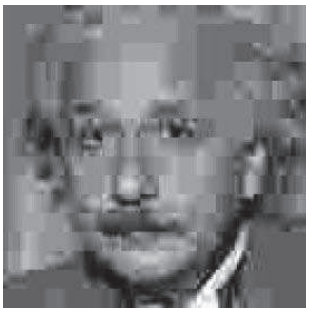
(b) PSNR=26.55



(c) PSNR=26.55



(d) PSNR=26.60



(e) PSNR=26.55

Figure 1.2: Einstein original image (a) and different distorted versions of it. The same PSNR but different perceptual quality. b) Mean Shifted Image, c) Contrast Stretched Image, d) Blurred Image and e) JPEG Compressed Image

In relation with human perception MSE and PSNR are widely criticized [13, 10, 14, 15]. PSNR do not always agree with the evaluations of the Human

Visual System (HVS), therefore when it is used to predict, or correlate results, with human perception of fidelity and quality, it seams not to be the best choice. The human eye, for example, does not observe small changes of intensity between individual pixels, but is sensitive to the changes in the average value and contrast in larger regions. Another deficiency of these distortion functions is that they measure only local, pixel-by-pixel differences, and do not consider global artifacts, such as blockiness, blurring, jaggedness of the edges, ringing or any other type of structural degradation of the image.

The visibility of distortions depends on the image background, a property known as masking (see section 1.3.8). Distortions are often much more disturbing in relatively smooth areas of an image than in texture regions with a lot of activity, an effect not taken into account by pixel based metrics. Therefore the perceived quality of images with the same PSNR can actually be very different. An illustrative example is shown in Figure 1.2 where an original is altered by different types of distortions. Note that the PSNR values, relative to the original image 1.2(a) of several distorted images are nearly identical, even though the images present dramatically and obvious different visual quality. In [12] the problem with MSE is deeply studied.

But they, are still the most widely used metrics in comparisons of encoder performance. This, as we will see later can produce erroneous conclusions about the goodness of a specific encoding proposal. Nevertheless, some authors [16] argue that in scenarios with fixed content distorted by typical compression and channel artifacts, PSNR predicts the perceived subjective quality nearly as well as more complex quality models representing the state-of-the-art.

The aim of research in the field of image and video objective quality assessment, is to design quality metrics that can automatically predict and rank the quality of an image or video sequence giving a quality value that is high correlated to the subjective MOS or DMOS value given by human observers. This metrics are valuable because they provide image and video encoder designers, and standards organizations, with means for making meaningful quality evaluations without convening viewer panels and provides big saving in time and effort.

So, one of the objective in this work is to find, among the most important image objective quality assessment metrics, one that exhibits a good behavior for a large set of image (or intra-mode encoded video) distortions providing measures as much as close to the ones perceived by human observers and fast enough for their practical use.

In the literature, there is a consensus in a primer classification of objective

quality metrics [17, 18, 10] attending to the availability of original non-distorted info (the reference) to measure the quality degradation of an available distorted version:

- Full Reference (FR) metrics perform the distortion measure with a full access to the original version which it is taken as a perfect reference.
- No Reference (NR) metrics have no access to reference. So, they have to perform the distortion estimation only from the distorted version. In general they have lower complexity but are less accurate than FR metrics and are designed for a limited set of distortions.
- Reduced Reference (RR) metrics work with some information about the reference (similar to a perceptual hash algorithm). A RR metric defines what information have to be extracted form the reference, so it can be compared with the same information extracted from the distorted version. This reference side information is the only information available to the metric to perform the quality assessment.

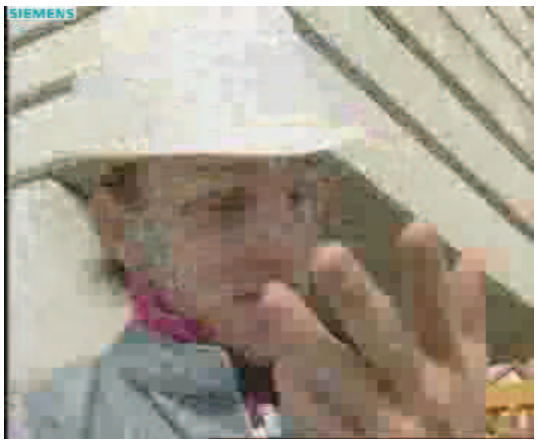


Figure 1.3: Artifacts: Blockiness

The most widely used FR objective video quality metrics by the scientific community, as mentioned before, are MSE and PSNR. In the last years, new objective image and video quality metrics have been proposed, mostly for FR/RR quality assessment . They emulate human perception of image/video quality since they produce results which are very similar to those obtained from subjective methods.

Most of these proposals were tested in the different phases carried out by the Video Quality Experts Group (VQEG) which was formed to develop, validate and standardize new objective measurement methods for video quality. The models that the VQEG forum validates result in International Telecommunication Union (ITU) recommendations and standards for objective quality models for both television and multimedia applications [19].

1.2 Principal coding artifacts and visual distortions

Figure 1.4: Artifacts: Blur

Most of the image or video compression algorithms used in the coding standards relay on the use of the DCT or the Wavelet transform. In such coding schemes the quality of the reconstructed version of the scene is deteriorated by the loss of information and by the introduction of coding artifacts. The loss of information is produced in the quantization step of the coding chain, while other artifacts can be introduced in other steps of the chain.

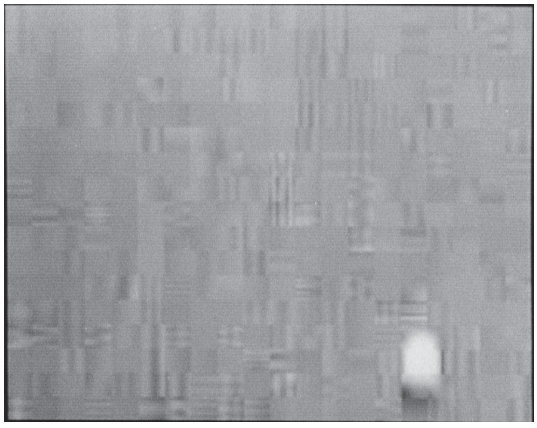


Figure 1.5: Artifacts: DCT basis image

Evaluation and classification of image coding artifacts [20] and video coding artifacts [21] is important in order to evaluate the performance of coding software and hardware products proliferating in the telecommunications, entertainment, multimedia and consumer electronics markets. A comprehensive classification will also assist in the design of more effective adaptive quantization algorithms

Figure 1.6: Artifacts: Ringing on DWT

and coding mechanisms in order to improve image and video codec performance. But, due to the complexity of the HVS, the perceived distortion is not directly proportional to the absolute quantization error [21].

In addition, our perceptual response to visual distortion, varies depending not only where quantization errors occur, but also how they coincide with structural image elements [22]. So, it is not possible to predict the quantization level or the bit-rate at which a specific artifact appears. And due to the different varieties of bit-allocation techniques that have been proposed, which may, or may not, exploit the masking effects of the HVS, this prediction is even more complicated.

Nevertheless, many efforts have been done to perform adaptive quantization to reduce artifacts produced by encoders that use specific transforms, like DCT [23, 24, 25, 26] and DWT [27, 28, 29]. In addition some specific artifacts produced by the DCT transform, like blocking, are eliminated by the use of DWT techinques [30].

The classification of coding artifacts is important too, in the design of filtering and for the search of objective psychovisual-based quality metrics.

Noise and artifact are terms used to describe speckles, spikes, missing data, and other marks, impairments, defects and abnormalities in image data created during the acquisition, transmission, and processing of image data.

The following, summarizes the most common noise and artifacts produced mainly in the processing of image data, describing only, how they manifest and possibles causes and relationships. Some of these effects arise only in block-based DCT schemes, others only in DWT schemes where the transform is applied to the whole image/frame, and finally some of them arise in block-based DWT schemes like JPEG2000. For example, in the LTW encoder [31], the transform is applied to the entire image, therefore none of the block-related artifacts occur. Instead, blurring and ringing are the most prominent distortions in these type of encoders. Figures 1.3 to 1.8 show some of these artifacts.

- The blocking effect or blockiness (figure 1.3), refers to the appearance of a block pattern in the reconstructed sequence. It is due to the independent quantization of individual blocks (usually of 8x8, 16x16, etc.. pixels in size) in block-based DCT coding schemes. It is more visible in low-detail regions when coarse quantization is applied to adjacent blocks, producing discontinuities at the boundaries of that blocks. The blocking effect is often the most prominent visual distortion in a compressed sequence due to the regularity and extent of the pattern. The false edges of the blocking effect are perceived as abnormal high frequency components in the spectrum of the image.

- Blurring manifests itself as a loss of spatial detail and a reduction of edge sharpness in regions with moderate and high detail (figure 1.4). Different types of blurring may occur. Motion blur due to the relative motion between elements in the scene, out focus blur (defocused camera or lens aberrations) and blur can be also introduced when compressing the image. It is due to filtering and the suppression of the high-frequency coefficients by coarse quantization i.e. an image appears blurred when its high spatial frequency in the spectrum is attenuated. Blurring means that the received image is smoother than the original.

- Color bleeding is the smearing of the color between areas of strongly differing chrominance, typically near edges over flat backgrounds. It results from the suppression of high-frequency coefficients of the chroma components.

- Each of the DCT basis images have a distinctive regular horizontally or

vertically oriented pattern which make them visually conspicuous (figure 1.5). The DCT basis image effect is prominent when a single DCT coefficient is dominant in a block. The effect is caused by coarse quantization of the AC DCT coefficients in areas of high spatial activity within a frame, resulting in the nullification of the low-magnitude DCT coefficients which are within the quantization dead-zone.

- Slanted lines often exhibit the staircase effect. It is due to the fact that DCT basis images are best suited to the representation of horizontal and vertical lines, whereas lines with other orientations require higher-frequency DCT coefficients for accurate reconstruction. The typically strong quantization of these coefficients causes slanted lines to appear jagged.

- Ringing artifacts manifest themselves in the form of ripples or oscillations around high-contrast edges in compressed images. They can range from imperceptible to very annoying, depending on the data source, target bit rate, or underlying compression scheme (figure 1.6). Ringing is fundamentally associated with Gibbs' phenomenon and is thus most evident along high-contrast edges in otherwise smooth areas. It is a direct result of improper quantization of high-frequency, leading to irregularities in the reconstruction. Ringing occurs with both luminance and chroma components. Since the high-frequency components play a significant role in the representation of an edge, coarse quantization in this frequency range (i.e., truncation of the high-frequency transform coefficients) consequently results in apparent irregularities around edges in the spatial domain, which are usually referred to as ringing artifacts.

- False edges are a consequence of the transfer of block-boundary discontinuities due to the blocking effect from reference frames into the predicted frame by motion compensation.

- Jagged motion can be due to poor performance of the motion estimation. Block-based motion estimation works best when the movement of all pixels in a macroblock is identical. When the residual error of motion prediction is large, it is coarsely quantized.

- Motion estimation is often conducted with the luminance component only, yet the same motion vector is used for the chroma components. This can result in chrominance mismatch for a macroblock.

- Mosquito noise is a temporal artifact seen mainly in smoothly textured regions as luminance/chrominance fluctuations around high-contrast edges or moving

objects. It is a consequence of the varied coding of the same area of a scene in consecutive frames of a sequence.

- Flickering appears when a scene has high texture content. Texture blocks are compressed with varying quantization factors over time, which results in a visible flickering effect while watching the sequence.
- Aliasing can be noticed when the content of the scene is above the Nyquist rate, either spatially or temporally.
- Masking is the reduction in the visibility of one component (the target) due to the presence of another (the masker). There are two kind of masking effects, luminance masking (light adaptation) and texture masking, which occurs when masker and target have similar frequencies and orientations.
- Jitter, in video sequences this distortion occurs due to abrupt variations resulting from asynchronous acquisition of video frames
- Jerkiness, refers to the perception of still images in a video sequence resulting from too low frame rates.
- Frame-loss is the loss of entire frames, normally frame-loss is produced in burst of different duration, i.e. number of frames. When frame-loss occurs, the video codec use to proceed by repeating the last correctly received frame (frame-freeze effect) or by setting a black frame. Frame-freeze is considered to be detected when its duration exceeds a certain threshold.

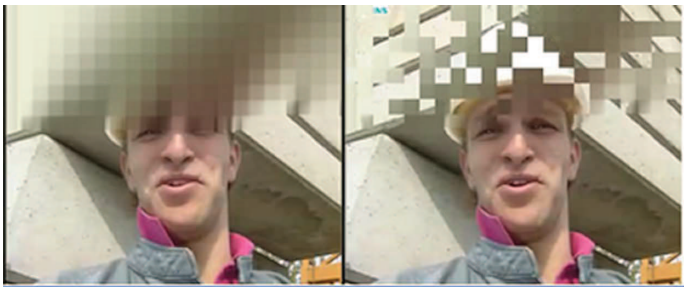


Figure 1.7: Artifacts: To types of reconstructed frames after packet losses

Another type of distortions are due to transmission errors of the bitstream over a noisy channel. When compressed video is transmitted over a packet-switched network, wired or wireless, some transport protocol like ATM (Asynchronous Transfer Mode) or the TCP/IP (Transfer Control Protocol / Internet Protocol) ensures the delivery of the bitstream. Normally the bitstream is

packetized, i.e. splitted in packets, whose headers contain sequencing and timing information. When the final application requires the bitstream in real-time for decoding and display the multimedia content, some common network conditions can produce the loss of some packets, which finally result in visual artifacts in the reconstructed sequence (figure 1.7).

In addition to the loss of packets, bit-errors can occur inside packets which are not lost, producing several type of noise effects in the reconstructed image or frame (figure 1.8), that are different depending on, the codec being use and many other factors as bits allocation in the bitstream, amount of bits (burst error), importance of the bits for the coding scheme, etc.

Packets can be lost or delayed, so that they are not received in time to be decoded when requested. To the decoder both alternatives have the same effect, the packet is lost and the bitstream can not be completely decoded. If some packets need dependent information contained in lost packet, for example, information that is differentially predicted, then the lost of a single packet corrupts the rest of the packets until the reception of the first non-dependent packet.

Figure 1.8: Artifacts: Bit Errors on DWT

For example, an MPEG macroblock that is damaged through the loss of packets corrupts all following macroblocks until an end of slice is encountered, where the decoder can resynchronize. In this example two types of errors are produced by the loss of packets, a spatial loss propagation and a temporal loss propagation. The spatial loss propagation is due to the fact that the DC

coefficient of a macroblock is differentially predicted between macroblocks. The temporal loss propagation arise when the lost information is needed by motion estimation.

The visual effect of such loss depend on the ability of the decoder to deal with corrupted bitstreams. Some decoders include clever concealment techniques, such as early synchronization and spatial or temporal interpolation in order to minimize these effects.

1.3 Brief Overview of HVS

Some brief introduction to the Human Visual System must be done in order to understand how the Objective Quality Assessment Metrics are build. Only the most important characteristics of the HVS that are implemented in these metrics are here briefly reviewed [17, 32, 33, 18, 10].

1.3.1 The Visual Pathway

The first contact of light wiht the eye is at the cornea, the main refractive surface of the eye, see Figure 1.9 from [18], then enters the eye through the pupil, in the center of the iris. The pupil diameter varies from 3 to 7 mm, and changes it size up to a factor of 5, based on the prevailing light level and other influences of the nervous system.

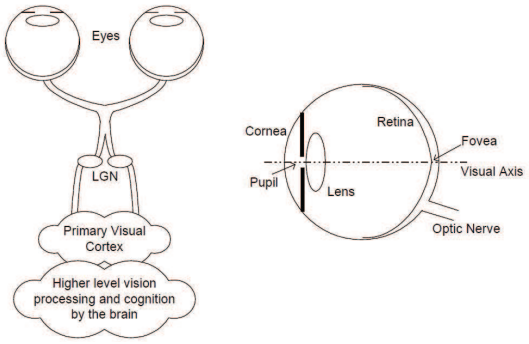


Figure 1.9: Schematic diagram of the human visual system

The light goes through the lens, that changes it shape with accommodation to focus the image on the back of the eye, projecting an inverted image of the visual field. After the lens, light passes through the gelatinous vitreous humor in the main body of the eye.

At the back of the eye is the retina, an extension of the central nervous system, where the light sensitive photoreceptors transduce the electromagnetic energy of light into the electro-chemical signals used by the nervous system. It consists of five main neural cell types organized into cellular layers and synaptic layers.

The photoreceptors, that initiate the neural response to light, are located on the outer part of the retina. There are two classes of photoreceptors, rods and cones. The rods are responsible for vision at very low light levels (scotopic) and do not normally contribute to color vision. The cones, which operate at higher light levels

(photopic), mediate color vision and the seeing of fine spatial detail, so they are responsible for vision in normal light conditions. There are three different types of cones, corresponding to three different light wavelengths. The L-cones, M-cones and S-cones (corresponding to the Long, Medium and Short wavelengths) split the image projected onto the retina into three visual streams. These visual streams can be thought of as the Red, Green and Blue color components of the visual stimulus, though the approximation is crude.

The photoreceptors are non uniformly distributed over the retina. The point on the retina that lies on the visual axis is called the fovea and it has the highest density of cone cells. This density falls off rapidly with distance from the fovea. The distribution of the ganglion cells, the neurons that carry the electrical signal from the eye to the brain through the optic nerve, is also highly non-uniform, and drops off even faster than the density of the cone receptors. The net effect is that the HVS cannot perceive the entire visual stimulus at uniform resolution.

The signals from the photoreceptors are processed via of retinal connections and exit the eye by way of the optic nerve. The axons of the ganglion cells, in the inner cellular layer of the retina, are gathered together and exit the eye at the optic disc forming the optic nerve that projects to the Lateral Geniculate Nucleus (LGN), a part of the thalamus in the midbrain. These synaptic connections to neurons, projects to the primary visual cortex which contains neurons tuned to various aspects of the incoming streams, such as spatial and temporal frequencies, orientations and directions of motion. These areas in the visual cortex respond to visual stimuli and processes of various modes of vision such as form, location, motion, color, etc.

The neurons in the cortex have receptive fields that are modeled as two-dimensional Gabor functions, which are linear filters that typically is used for edge detection. The whole set of these neurons are modeled as an octave-band Gabor filter bank [34] where the spatial frequency spectrum (in polar representation) is sampled at octave intervals in the radial frequency dimension and uniform intervals in the orientation dimension. The output of these neurons saturates as the input contrast increases. The tasks of these neurons in the cortex, is typically emulated in some quality assessment metrics and perceptually driven encoders, with the inclusion of models of spatial frequency and orientation selectivity.

1.3.2 Foveal and Peripheral Vision

As stated before, the retinal image is a distorted version of the input visual field. A natural noticeable distortion is blurring, produced by imperfections of the optics of

the eye and natural variations of light produced at each step in the visual pathway.

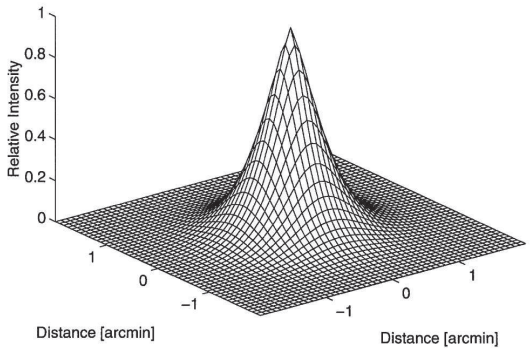


Figure 1.10: Point spread function of the human eye as function of visual angle

To quantify and model the amount of blurring of a HVS a Point Spread Function (PSF) or a Line Spread Function (LSF) is used. Its Fourier transform is the Modulation Transfer Function (MTF) of the eye for this stimulus. The amount of spreading or blurring of a stimulus is a measure of the quality of an optic system. The amount of blurring depends on the pupil size being higher as the pupil increases it size due to lower ambient light intensities.

This variation is modeled by a a simple formula (Equation 1.1 [17]) to approximate the foveal point spread function of the human eye with good focus and a pupil diameter of 3 mm. [35], being α minutes of arc. This PSF, presented in Figure 1.10, also changes with wavelength. By accommodation, the eye can place any wavelength into good focus, but it is impossible to focus all wavelengths simultaneously.

$$PSF(\alpha) = 0.952e^{-2.59|\alpha|^{1.36}} + 0.048e^{-2.43|\alpha|^{1.74}} \tag{1.1}$$

As commented in section 1.3.1 the densities of the cone cells and the ganglion cells in the retina is not uniform. The number of photoreceptors have a peak at the fovea and decreases with distance from it. Cones are concentrated in the fovea, the region of highest visual acuity, which covers approximately two degrees of visual angle on the retina. When a human observer fixates at a point of the visual scene, this point is located at the fovea being sampled with the highest spatial resolution. The surrounding points of the scene are progressively processed with lower spatial resolutions. The high-resolution vision due to fixation by the observer onto a region is called foveal vision, while the progressively lower resolution vision is called peripheral vision.

Regarding the visual spatial acuity of the fovea, the photoreceptors are packed tighgly in triangular arrangment with a mean center-to-center spacing of 32 arc min. [36] This corresponds to a sampling rate of approximately 120 samples per optical degree or a Nyquist frequency of around 60 cpd (cycles per optical degree). Visual spatial acuity is therefore considered to be approximately 60 cpd although under special conditions, for example, peripheral vision and large pupil sizes higher spatial frequencies can be either directly resolved.

Image quality assessment models [37, 38, 39] can include foveal vision in its implementation. These models also introduce vision modeling taking into account the non-uniform distribution off cones in the retina, modeling the image with less resolution as the distance from the region of interest (foveated part of the image) increases. Foveal vision models can resample the image with the same density of the receptors in the fovea in order to provide a better approximation of the HVS.

Most models neglect eccentricity and off-axis effects and concentrate their modeling efforts on the properties of the fovea. This is usually justified with the fact that when the eyes bring into the fovea part of the image, this part is sampled at highest resolution, being any part of the image processed in the same way. As the the optical and retinal properties are relatively uniform across the fovea, using the same properties for the whole image significantly simplifies modeling.

1.3.3 Contrast Sensitivity

As commented in section 1.3.5 the HVS can perceive small differences in luminance. However the minimal difference that still can be perceived, depends on the background luminance. The dependence to the background luminance that the HVS has while detecting differences in the luminance is called Contrast Sensitivity. That is, sensitivity to intensity differences, is dependent on the local luminance in regions of the image [40]. A basic model for this dependence is the Weber-Fechner law. It states that, sensitivity to luminance differences in a stimulus is proportional to the mean luminance of the stimulus. Mathematically, Weber contrast can be expressed as Equation 1.2

$$C^W = \frac{\Delta L}{L} \tag{1.2}$$

The Weber-Fechner law is not fulfilled for all background luminance levels. It holds for luminance levels above approximately $10 cd/m^2$ [41], below this level the contrast threshold increases as luminance decreases, i.e. there is less sensitivity to contrast below this level. Evidently, the Weber-Fechner law is only

an approximation of the actual sensory perception, but contrast measures based on this concept are widely used in vision science.

Contrast is the difference in the luminance level of adjacent parts of an image or visual field. That is, contrast is the difference in luminance or color that makes an object distinguishable. HVS is more sensitive to luminance changes (contrast) than to absolute luminance, so we can perceive objects regardless of the changes in illumination (above $10 cd/m^2$ as Weber-Fechner law states) as long as the contrast is high enough.

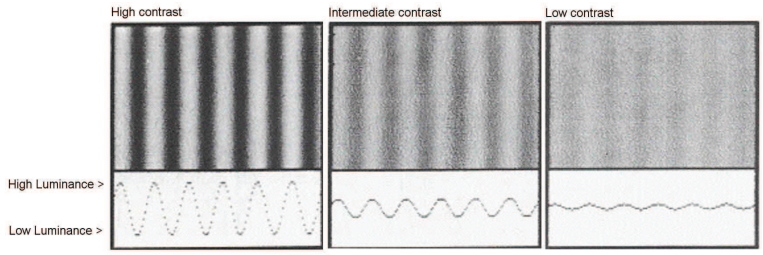


Figure 1.11: Three sine wave gratings with the same spatial frequency but with descending contrast from left to right

If contrast is too low we can not distinguish an object from the background. In this situation some objects in the scene turn into invisible objects. These objects are said to be below the contrast threshold.

The sensitivity is the inverse of the contrast threshold, i.e. $Sensitivity = 1/threshold$. Therefore, the smaller the contrast we need to perceive an object in the scene is, the higher is our sensitivity. And the opposite, for low sensitivity we need higher contrast to perceive differences. Under optimal conditions, the contrast threshold can be less than 1%.

Suppose a scene where the contrast of an object with its background is descending, then at just the point where the object becomes invisible we could record the value of the difference in luminance between the object the background, this value is our contrast threshold. Its inverse is our contrast sensitivity. For example, if contrast threshold is 0.1 then sensitivity is $1/0.1 = 10$, if threshold is 0.01 then sensitivity is 100, and so on.

In Figure 1.11 we can see three gratings, these gratings are called sinusoidal gratings or sine wave gratings, because they change gradually in luminance over space (horizontal axis). At the bottom of each grating a sine wave represents the luminance variability in the horizontal axis.

The contrast of periodic (often sinusoidal) stimuli with varying frequencies is

defined by the Michelson contrast. The Michelson definition of contrast is in fact $(LMAX - LMIN)/(LMAX + LMIN)$ where $LMAX$ and $LMIN$ stands for Max Luminance and Min Luminance respectively. If the sine wave of the rightmost grating in Figure 1.11 were just a horizontal line there would be no contrast at all, then, the so-called grating would just be a homogeneous gray, $LMAX$ would be the same as $LMIN$ and contrast would be zero because $(LMAX - LMIN)$ would be zero. If, on the other hand, the black bars were very black and the white bars were very white, $(LMAX - LMIN)/(LMAX + LMIN)$ might be $(1000 - 1)/(1000 + 1)$, so the maximum contrast you can ever have is 1.0

But, if in the previous scene are more than one object and these objects are quite different in size, shape and texture, then the point in which each object becomes invisible is different. This is due to the fact that the human perception off contrast not only depends on the difference of luminance but also on the spatial frequency. So, the contrast threshold varies with the spatial frequency.

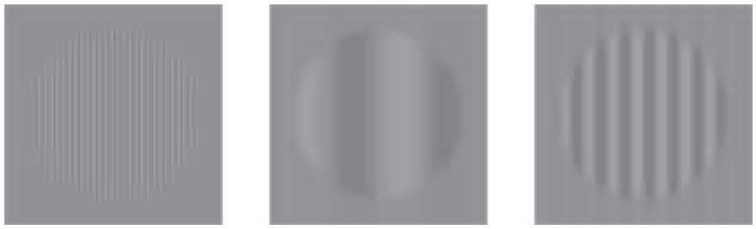


Figure 1.12: Which of these three gratings appears highest in contrast and which appears lowest in contrast?

In [42] we can find a very clear explanation of contrast sensitivity. To illustrate this we can see figure 1.12 from [42] where three gratings are presented. Most people would rank them in the order shown, with the leftmost grating being the one with lower contrast. But this is wrong because all three gratings have precisely the same physical contrast.

Suppose we use a lens to cast an image of a target grating on a white paper. This target grating has a specific physical contrast that we call “target contrast”. Then, using a photometer we determine the intensity of the light and dark portions in the image and, hence, the contrast of the image of the grating produced by the lens, the “image measured contrast”. We repeat these measurements for different spatial frequencies always with gratings of the same “target contrast”.

If we graph the results, being the horizontal axis the spatial frequency of the grating, and the vertical axis the “image measured contrast” as percentage of the “target contrast”, then we get the a transfer function of how contrast is transferred through the lens, see Figure 1.13. In this figure two curves appear, one for a clean

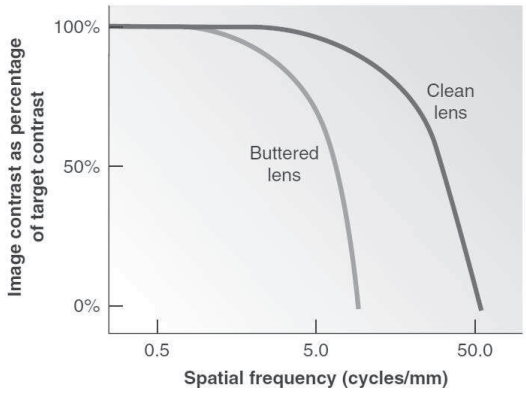


Figure 1.13: Two transfer functions for a lens. How contrast in the image formed by the lens is related to contrast in the object.

lens and another corresponding to a buttered lens, i.e. smeared with a buttery finger.

For the clean lens curve, up to a specific spatial frequency the contrast in the image is identical to that of the target, but for higher frequencies the lens reproduces the target less faithfully. The frequency at which the contrast falls to zero is called the cutoff frequency, when the frequency exceeds this value the image and the target (if a perfect lens) will no longer contain any contrast.

The curve for the buttered lens, has a lower cutoff frequency, degrading the contrast of the target more rapidly than the clean lens. But at very low frequencies the smear makes little difference in the performance of the lens. This means that a high quality lens reproduce better fine and coarse spatial detail whereas a low-quality lens only reproduce well low frequencies. Think about when you are wearing smeared glasses.

Natural scenes are not as simple as gratings and that images are composed of many different spatial frequencies, sine waves in any orientation. We can treat the scene as a sum of a series of simple sinusoidal components, by using Fourier analysis, we can evaluate how the lens reproduce each of those components. So we can first determine the transfer function of the lens (supose the buttered one) and second analyze the visual scene into its spatial frequency components. Finally, with this information we can conclude which spatial frequency components will be preserved by the lens in the image and which will not.

Suppose now that the lens is our Human Visual System, which frequencies will we perceive an which one not? The problem here is that is not as easy as in the case of the lens, to determine the transfer function of our HVS.

1.3.4 The Contrast Sensitivity Function

With the HVS we can not reproduce the procedure employed with the lens in order to measure the frequency components of the gratings that are preserved in the image, because the image is formed inside the eye. Moreover, this image would give information of only a part of the complete transfer function of the HVS, because other neural and cognitive components of it, further processes that image.

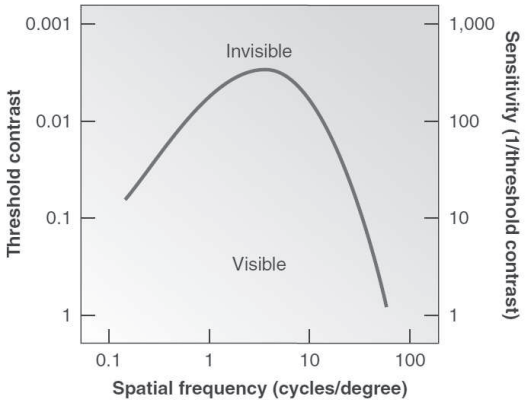


Figure 1.14: Contrast sensitivity function shape.

As we are interested in visual perception we must be concerned with the perceptual transfer function which depends on the optical transfer function and the neural and cognitive transfer functions. By measuring contrast thresholds for different spatial frequency gratings, we can derive a curve that describes the entire visual system's sensitivity to contrast. We call this curve the Contrast Sensitivity Function (CSF), to distinguish it from the transfer function of a lens.

Figure 1.14 shows the CSF for a human adult. The horizontal axis specifies the spatial frequency plotted as the number of cycles within a degree of visual angle. The vertical axes plot the minimum contrast required to see the grating where left axis show units of contrast and right axis inverse of this contrast value (defined as sensitivity). This curve defines the window of visibility, that is, underneath the curve represents combinations of contrast and spatial frequency that can be seen, while above represents combinations that can not be seen.

The CSF curve in figure 1.14 differs from the lens transfer functions of Figure 1.13 at low frequencies because the HVS is less sensitive to very low spatial frequencies than it is to intermediate ones. Objects of a visual scene which have most of their spatial frequency information around the optimum

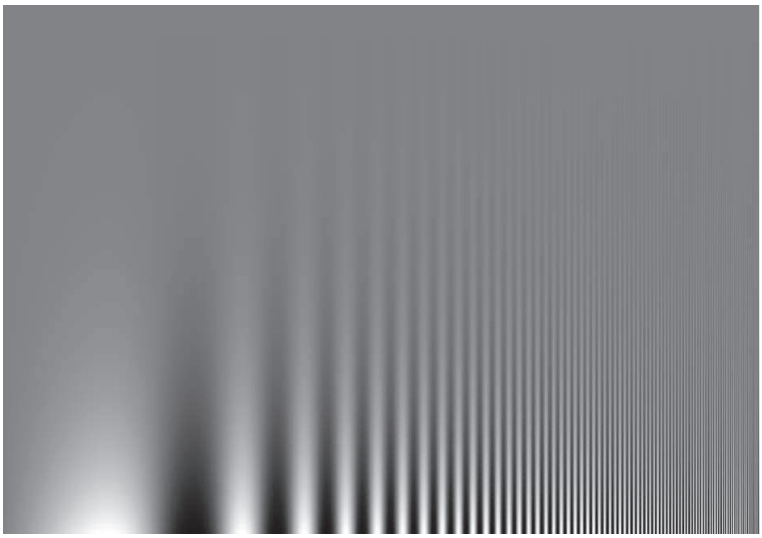


Figure 1.15: Campbell-Robson contrast sensitivity chart

point on the CSF will be clearly visible even when they are in low contrast. But if these objects have very low spatial frequencies (very large objects) or only very high spatial frequencies (very small objects or very fine details of them) they will be less visible and their contrast should be higher in order to be seen. This explains why the gratins in figure 1.12 appear different in contrast: their apparent contrast varies with your sensitivity to different spatial frequencies.

Figure 1.15, the so-called Campbell-Robson chart [43] demonstrates the shape of the spatial CSF for sinusoidal stimuli in a very intuitive manner. The luminance of pixels is modulated sinusoidally along the horizontal dimension. The frequency of modulation increases exponentially from left to right, while the contrast decreases exponentially from 100% to about 0.5% from bottom to top. The minimum and maximum luminance remain constant along a given horizontal line through the image. The location of its peak depends on the viewing distance.

Campbell [44] suggested that the CSF does not reflect the sensitivity of a single mechanism, but the combined activity of sets of neurons, each capable of responding to targets over only a restricted range of spatial frequencies. This independent mechanisms, called 'filters', 'detectors' or 'channels' are responsive for detecting luminance variations that occur at a particular spatial scale (frequency). Some respond to the coarse variations and others to finer details. So, the CSF reflects the envelope of sensitivities of multiple filters see figure 1.16. Consequently the HVS uses the spatial frequency filters to perform a type of Fourier analysis of the retinal image.

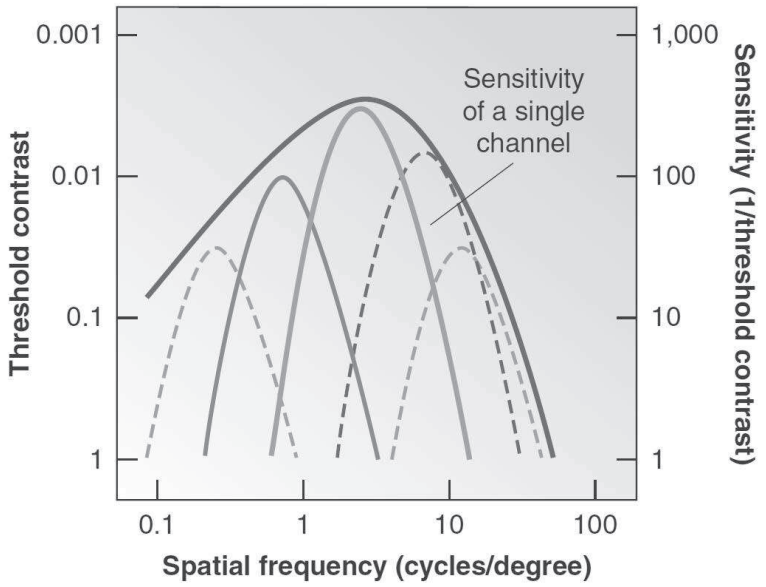


Figure 1.16: Multiple filters CSF model.

1.3.5 CSF and light conditions

The HVS operates over a wide range of light intensity values. The scotopic and photopic vision cover actually 12 orders of magnitude, varying from the detection of a single photon to extremely bright day-light conditions. To reach this dynamic range more than a single adaptation process is involved. The first adaptation mechanism is located in the pupil, which resizing mechanism controls the amount of light entering the eye. Then, a more powerful regulatory process of light adaptation is hold in the photoreceptors and other retinal cells adjusting the gain of post-receptor neurons in the retina. The retina encodes the contrast of the visual stimulus instead of coding absolute light intensities. There are two different adaptation processes:

- Light adaptation. This adaptation happens very quickly. Sensitivity changes from dark light to bright light conditions. A decrease of the chemical concentration in the photoreceptors is the cause.

- Dark adaptation. Adaptation from bright light into darkness. In this case the chemical concentrations increases, but this process is very slow in comparison with light adaptation, it can take up to an hour until the chemical concentrations reaches its final state.

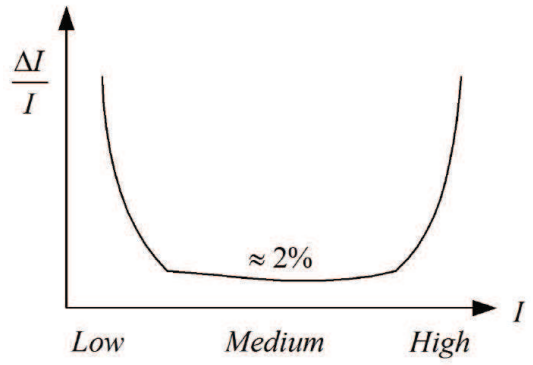


Figure 1.17: Contrast ratio: Weber fraction

The response of the eye to changes in the intensity of illumination is nonlinear. If we consider a patch of light intensity surrounded by a background intensity I, we can define as Just Noticeable Difference (JND) as the smallest increment ΔI in luminance perceived by our HVS, [45] states that the sensitivity of human eyes to discriminate these increments depends not only on the difference itself but also on the level of intensity. Over a wide range of intensities the Weber fraction $\frac{\Delta I}{I}$ is nearly constant at a value of about 0.02, but this result does not hold for very low or very high light intensities as shown in figure 1.17 where $\frac{I+\Delta I}{I}$ represents the contrast ratio. So, the Contrast Sensitivity is also affected by the luminance level.

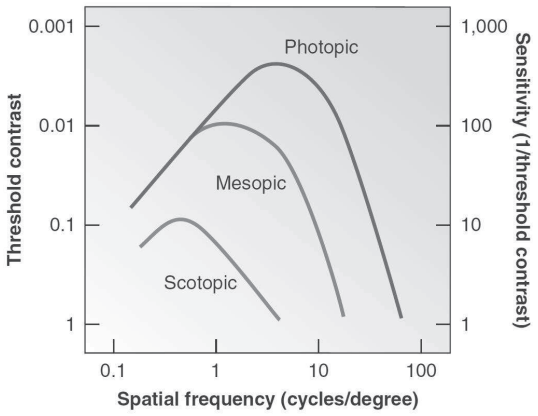


Figure 1.18: CSF under different luminance conditions

Figure 1.18 depicts how the CSF varies with light conditions showing three CSF curves, the photopic curve (datytime), the mesopic curve (twilight) and

scotopic curve (dim light). As the level of light decreases from daylight to twilight, visual sensitivity drops primarily at high spatial frequencies, that is why is difficult to read small letters (small details) in twilight, and lower frequencies are little affected. When light drops further, sensitivity decreases even at low frequencies.

1.3.6 Chromatic CSF

Contrast sensitivity to chromatic spatial variations has also been studied [46] using harmonic stimuli, measuring red-green and blue-yellow gratings. Figure 1.19 from [10] shows the chromatic CSF curves in addition to luminance CSF curve. The color CSFs are characterized as a low-pass filter with high frequencies cut-off at much lower frequencies than the cut-off for luminance curve. That studies reveals that the acuity of the blue-yellow channel is limited by the distribution of the S-cones in the retina, but the red-green channel is limited by subsequent neural processing.

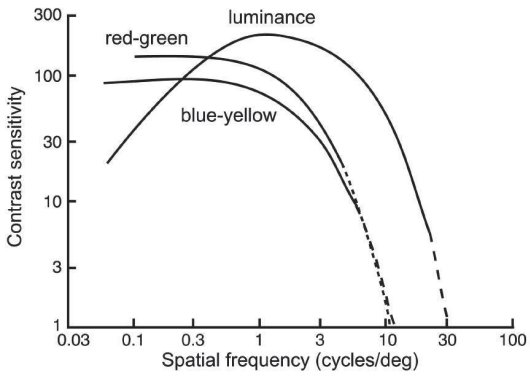


Figure 1.19: Contrast Sensitivity Functions of chromatic and luminance components

The sharpness of an image is judged based on the sharpness of the luminance information since the visual system is not able of solve high-frequency chromatic information . This fact has been used in the compression and transmission of color images since high frequency chromatic information can be removed without a loss in perceived image quality [46, 10]. The full range of colors is perceived only at low frequencies [47].

1.3.7 Temporal CSF

The human contrast sensitivity depends on the color, the spatial and also on the temporal frequency of the stimuli. Similar as the spatial CSF, the temporal CSF also has a low-pass behavior. The interaction between spatial and temporal frequencies are commonly used in vision models for video [48].

The spatio-temporal CSF approximations [47] are shown in figure 1.20. Achromatic spatio-temporal contrast sensitivity is higher than chromatic sensitivity, especially for medium-high spatio-temporal frequencies. In the achromatic chart of figure 1.20 we can see that for low spatio-temporal frequencies our sensitivity decreases whereas chromatic sensitivity does not. As stated before, the full range of colors are perceived at low frequencies, spatial and temporal frequencies as shown in the chromatic chart of figure 1.20. At higher frequencies sensitivity to blue-yellow frequencies declines first and at even higher frequencies sensitivity to red-green stimuli declines too and perception becomes achromatic [47].

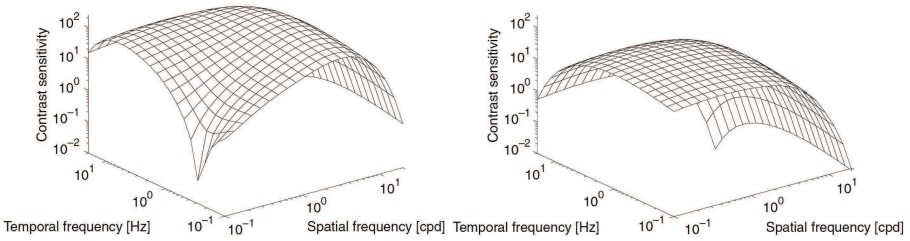


Figure 1.20: Approximations of the achromatic CSF (left) and the Chromatic CSF (right)

In has been some controversial in the literature about the space-time separability of the spatio-temporal CSF. From a modeling and usability point of view, separability is a very interesting property in order to process video in such a way that takes into account the temporal dimension of the HVS sensitivity to contrast.

Early studies conclude that the spatio-temporal CSF is not space-time separable at lower frequencies [49, 50]. Further studies [51, 52] conclude that spatio-temporal CSF can be approximated by combinations of separable components in space and time. And again later studies confirm the inseparability of space-time dimensions in the spatio-temporarl CSF [53].

1.3.8 Masking

Masking is an important phenomenon in vision as it reflects the relationships and interactions between different stimuli. It occurs when a stimuli, that is visible by itself, in the presence of another stimuli becomes invisible.

There is a relationship between both stimuli, the masker and the original stimuli. Some similar characteristics in both stimuli causes the invisibility of the original stimuli when the masker is present, normally this interaction occurs gradually as these related properties change. These properties are the spatial frequency, the orientation and the phase of the masker relative to the original stimuli, i.e. the masking effect is maximum when the stimulus and the masker are closely coupled in terms of orientation, spatial and temporal frequency, and decreases rapidly as the distance between the signals increase in the spectral domain.

Sometimes the opposite effect occurs, facilitation, when a stimuli cause the perception of another stimuli that was not perceived before.

When talking about quality assessment, normally is helpful to think that the distortions produced by compression, transmission, coding noise or whatever other artifacts (original stimuli) are masked or facilitated by the image or sequence being compressed, transmitted or coded, that acts as background.

Spatial masking is strongest when the interacting stimuli have similar characteristics, i.e. similar frequencies, orientations, colors, etc. But it also occurs between stimuli of different orientation and between stimuli of different spatial frequency.

For example, in some regions of the image some noise or compression artifacts are more visible than in other parts, in that cases the background image is acting as masker for the artifacts, see figure 1.21 from [47] as example. The noise pattern in the top part of right image is also present in the bottom part of the same image, but the image content in this area, rocks and see, mask the noise.

So, it is important to understand which are the properties of both parts, the image in those regions and the noise or artifact itself, because this knowledge can lead to adaptive techniques to code, compress or transmit images in different ways at different regions.

Temporal masking accounts for the elevation of the visibility threshold due to temporal discontinuities in intensity. For example in transitions from dark to bright the threshold elevation may last up a few hundred milliseconds after transition.

Figure 1.21: The background image is acting as masker of a noise pattern. Left is the original image. In the right image the noise pattern is applied to the top and to the bottom of the image. The texture in watter and rocks makes difficult to detect the noise pattern.

Pattern adaptation is another type of masking that affect to the contrast sensitivity due to an adjustment of the visual system sensitivity in response to a prevalent stimulation pattern [47]. Adaptation of a certain spatial frequency can lead to noticeable decrease of contrast sensitivity around that frequency.

1.3.9 Suprahtreshold Contrast Sensitivity

Up to know, discussion was centered in at threshold sensitivity, i.e. our sensitivity at threshold level. Our sensitivity at threshold is very dependent on spatial frequencies, as shown in previous sections, i.e. it depends on the spatial frequency, and thus the contrast threshold varies, having a maximum sensitivity (lower contrast threshold) in the range from 2 to 6 cpd, and as said in section 1.3.5 this varies with luminance conditions too.

When we talk about suprathreshold sensitivity we are focusing in the visible

area of the CSF (see figure 1.14) which is the area of our regular visual conditions. There, the contrast level is above the threshold level, in other words, contrast is above the minimum level required for detect the target over the background.

The relationship between the perception of contrast and spatial frequency at levels above threshold is slightly different than at threshold. The effects perceived at threshold are qualitatively different from those at suprathreshold levels, so, models of detection and discrimination levels may not be applicable, because a "contrast constancy" effect (the apparent contrast matches physical contrast by an intra-channel response-gain control mechanism of the spatial frequency channels), is produced in the range from 1 to 10 cpd of spatial frequency [54, 55, 10].

The "contrast constancy" property [54] suggest that at suprathreshold levels the contrast ratios specified by the CSF would fail to indicate veridical measures of perceived contrast; rather, perceived contrast can be predicted based primarily on physical contrast.

The "contrast cosntancy" property and the effect that natural images, as masker, produce in the perception of suprathreshold targets was studied in [56] where experiments conclude that "contrast constancy" occurs only after an adaptation process and that natural images decrease the perceived contrast only of lower-frequency distortions.

In the context of lossy image compression, this "contrast constancy" property suggest that the contrasts of the distortions could be theoretically proportioned equally across the frequency spectrum (e.g., by assigning all frequency subbands equal weights) without affecting the total perceived contrast.

Because compression induced distortions are presented against a natural image maskers, then, under "contrast constancy" assumption and with the support of results [56] of authors experiments, it is reasonable to assume that the post-adaptation might also affect the perceived contrast of suprathreshold distortions in a similar fashion, and as natural images decrease the perceived contrast only of lower-frequency distortions, more contrast would be allocated to these lower-frequency distortions, e.g., by assigning the corresponding subbands smaller weights (indicating less "visual importance"). Experiments in the context of lossy image compression using the wavelet transform [57] confirm too that when distortions are suprathreshold, physical contrast is a better indicator of perceived contrast than predictions based on the CSF.

Authors in [57] detected also that although "contrast constancy" is observed too for wavelet subband quantization distortions at suprathreshold levels in their unmasked experiments (without natural-images as masker), when using

natural-images as masker, selective effects on the perceived contrast of low-frequency distortions are observed. Authors conclude that proportioning the contrast of the distortions according to the perceived contrast ratios, produce lower visual image quality that the one obtained by proportioning the contrast using CSF derived ratios. Authors also provide an explanation to this fact based on the global precedence mechanism, which sanctions the allocation of less contrast to lower-frequency distortion in order to preserve the visual integration of image features across scale-space.

Also in in Part I of the DWT based compression standard, JPEG 2000, the "contrast constancy" property is not applied and by the way of a visual progressive weighting factor, greater contrast allocation is given to higher-frequency distortions.

1.4 Objective quality assessment metrics

An objective quality assessment metric for images or video sequences, measures the perceived distortion of the image or the sequences without human intervention in such a way that results are highly correlate to the human quality ratings for the image or sequence. It can be use as part of a quality of service monitoring application to identify changes of quality over time or as part of a rate-distortion framework that seeks to optimize the quality of compressed images or sequences by minimizing the perceived distortion.

When comparing the performance of different image and video coding approaches, improvements of theses approaches or completely new codec designs, the most common way of doing the comparison between proposals, is in terms of the Rate/Distortion (R/D) behavior of the compared approaches. When using R/D comparisons, usually the distortion is measured in terms of PSNR (Peak Signal-to-Noise Ratio) values, while rates are often measured in bpp (bits per pixel) when comparing images or Kb/s (Kilobits per second) when comparing video sequences. However, it is well known that the PSNR metric not always capture the distortion perceived by the human being, see section 1.1.

So, a lot efforts were performed to define objective image and video quality metrics that are able to measure quality distortion closer to the one perceived by the destination user. In this section, we perform a study of different available objective image quality metrics in order to evaluate their behavior, taking as reference the classical PSNR metric. Our purpose is to find an image quality metric that is able to substitute PSNR for image quality assessment and video quality assessment in intra mode, and substitute the PSNR as distortion metric in the R/D comparisons with that metric, obtaining so, a perceptually more accurate R/D comparison when designing and evaluating image and video codec proposals.

The main objectives of using QAMs (Quality Assessment Metric) is to avoid the need to run MOS test and getting the most accurate perceptual quality value of images or video sequences. An objective QAM is told to have better behavior than other if its output quality values are best correlated with the quality values given by human observers, i.e. as close as possible to the quality perceived by humans, when a MOS test is performed. Metrics for assessing how good this correlation is are reviewed later in this section. So, QAM refers to the metrics and models for predicting this subjective visual quality scores, MOS or DMOS.

As summarized in section 1.2 many different types of distortions arise when processing, transmitting, encoding and compressing images or videos. An ideal objective quality metrics should exhibit a good behavior regardless of what kind

of distortions are affecting the image. Also it would be desirable that the time required for giving the quality measure is short enough for a practical use.

In the past years a big effort has been done in the field of QAM. A large number or metrics can be found in the literature. Some of them have been designed for a specific kind of distortions, while others are more generalist and try to perform regardless of the distortion type. Besides, each metric design is different. We provide a classification of image QAM. Objective evaluation of picture quality in line with human perception is still difficult [40, 13, 58, 59, 2, 60, 61] due to the complex, multidisciplinary nature of the problem, including aspects related to physiology, psychology, vision research and computer science. Nevertheless, with proper modeling of major underlying physiological and psychological phenomena, obtaining results from psychophysical tests and experiments, it is possible to develop better visual quality metrics to replace non-perceptual criteria as PSNR or MSE.

As mentioned in section 1.1 there is a consensus in a primer classification of objective quality metrics as Full Reference, No Reference and Reduced Reference. Most of the recently proposed image and video QAMs are Full Reference. They emulate and try to substitute the way in that human perception of image and video quality is use to score the perceived quality, in the sense that they produce results which are very similar to those obtained from subjective methods. Most of the FR metrics can also provide a spatial distortion or error map for each frame or, for video sequences where they provide a time series of frame level distortion scores.

The time needed to access in FR mode to both sequences is affordable for compression frameworks or applications that are not executed in real time, but not for real-time quality monitoring applications. In theses cases NR or RR metrics are used instead. They detect classes of artifacts or error patterns in images or sequences, as blocking or blurring, but distortions for which these metric have not been designed for, remain invisible. Therefore, although most RR metrics extract features from the original image that will be compared to the same features extracted form the distorted version, there are also some RR metrics that works as FR metrics but with reduced version of the original sequence. This is the case of the metric in [62, 63] that uses a low-bandwidth version of the reference for comparing with the low-bandwidth version of the distorted sequence.

VQEG provide a forum where algorithm developers and industry users meet to plan and execute validation tests of objective perceptual quality metrics. VQEG testing includes several subjective databases whose results are to be predicted by the objective video quality models under examination. The format of the source content, the nature of the degradations, the statistical techniques

and almost every aspect related to how to prepare the visual content and how to measure the results are parametrized and proposed by the VQEG. As to the design of each metric provide different output quality scales, the VQEG also proposes the method to compare those heterogeneous metrics by translating the results in their own scores into a common scale to make them comparable. Once a validation test has been completed, VQEG submits a final report to the ITU, which is ultimately responsible for preparing new standards for objective perceptual quality measurement.

VQEG has completed three validation tests. The first two tests, called VQEG Full-Reference Television Phase I (FRTV-I) [58] and Phase II (FRTV-II), covered quality measurement of standard definition television services using Full Reference models. The first test, FRTV-I , was completed in 2000. None of the models tested outperformed the PSNR. Accordingly, the initial standard, published by ITU-T Study Group 9 as Recommendation J.144 , included only informative appendices detailing objective models. The second test, FRTV-II, was completed in 2003 [59]. At the end of this validation effort, the ITU-T published an updated version of Recommendation J.144 [64] in which four objective models were included as standardized objective perceptual quality measurement methods. The third and most recent validation effort was aimed at evaluating objective perceptual quality models suitable for digital video quality measurement in multimedia applications. This project, VQEG Multimedia Phase I (MM-I), was completed in 2008 [65], and ITU-T Study Group 9 has subsequently published two new standards based on that report: ITU-T Recommendation J.247 [66] defines four new full-reference objective quality methods for multimedia, and ITU-T Recommendation J.246 [67] defines one new reduced-reference objective quality measurement method for multimedia.

1.4.1 Frameworks

QAM can be classified by many factors as, the metric architecture (number and type of blocks, stages or algorithms used in the metric design), the primary domain (space or frequency) where they work, the inclusion or not of HVS characteristics or HVS models in their design, and so on.

We have found in the literature different QAM reviews and different classifications [32, 47, 10, 68, 69, 15, 70, 71], but without finding a common consensus on how to fully classify them. Some of these reviews explain with great detail most of the metrics cited here, so only the main characteristics or most relevant aspects of the metrics will be exposed here.

We grouped QAM into three different frameworks depending on the way they

are designed and if its design is driven or not by any of the available HVS models.

- HVS Model Based Framework
- HVS Properties Framework
- Statistics of Natural Images Framework

If the design of one metric is not clearly based on any specific HVS model, then we move this metric out of the group of HVS modeled metrics. However, that metric can still use, somehow, one or more of the previously described HVS characteristics. The third framework is related to the statistic analysis and properties of the natural scenes.

So, in this section we will briefly describe the main ideas behind the different frameworks and the most relevant and cited QAM of each one. Normally that main ideas are translated to functional steps or computational phases that conform the metric architecture. For each of the frameworks we will explain briefly this phases or steps.

HVS Model Based Framework

A basic idea of any metric based on a HVS model is that subjective differences between two images can not be extracted from the given images (original and distorted one), but from their perceived version. As it is known the HVS produces several visual scene information reductions, carried out in different steps. The way in which this information reduction process of our HVS is modeled, is the key to obtain a good subjective fidelity metric.

This framework includes the metrics that are clearly based on a HVS model, i.e. their design follow the stages of any of the available HVS models. We include here the Error Sensitivity framework (ESF) [2], and also some other RR and NR metrics that are based on HVS models.

The "'Error Sensitivity Framework"' include mainly FR metrics based on HVS models, being a common stage in all of them the quantification of the strength of the errors between the reference and the distorted signals in a perceptually meaningful way, i.e. using the HVS model. Therefore practically all the metrics in this framework (Error Sensitivity) are FR.

Generally, the emulation of HVS is a bottom-up approach that follows the first retina processing stages to continue with different models about the visual cortex behavior, modeled as consecutive processing stages. Also, some metrics

deal with cognitive issues about the human visual processing modeling that issues as additional stages.

The main difference between the FR metrics of this framework is related with the way they perform the subband decomposition inspired in the complex HVS models [72, 73, 74], low cost decompositions in DCT [75, 76] or Wavelet [63] domains, and with other HVS related issues like in [77] where foveal vision is also taken into account and in [78] where focus of attention is considered. It is worth noting that a big percentage of proposed FR quality assessment models share the common error sensitivity based philosophy, see figure 1.22, which is motivated from psychophysical vision science research [18].

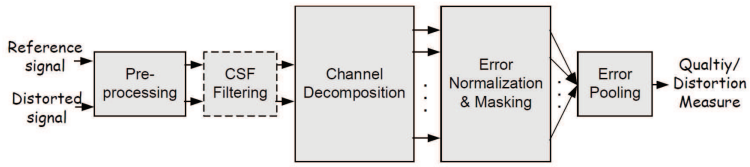


Figure 1.22: Common block diagram of the Error Sensitivity Framework

After some pre-processing in the space domain, usually the HVS models first decompose the input signal into spatio-temporal subbands in both the reference and distorted signal. As mentioned, this frequency decomposition is one of the biggest differences between models, and hence between metrics. Then, an error normalization, weighting process and masking process is carried out in order to give the estimated degradation measure.

Pre-Processing

In this stage, some pre-processing operations are done in order to adequate some characteristic of the reference and the distorted input versions. This operations commonly include pixel alignment, image cropping, color space transformations, device calibrations, PSF filtering, light adaptation, and other operations. Not all the metrics perform all this operations, each metric adjust the inputs in a different way.

A point to point misalignment can occur due to different reasons in the compression, processing and/or transmission of the reference image, so some metrics perform first a point to point correspondence that helps in upcoming stages to minimize assessment errors due to this fact.

Image crop are use by some authors [76, 74, 1, 79] in order to center processing in a region of interest or to avoid problems that arise in filtering stages

with image boundaries. Some authors perform also some segmentation process, in order to narrow the application scope of the metric to focus in these areas. In [1] a segmentation process is done in order to determine which are the "'dominant blocking areas"' based on the evidence that blocking artifacts are not noticeable likewise in all regions of the image.

Some metrics decide to convert the color signal to a space color that is better correlated to HVS. Author in [79] present a FR metric for color video sequences, based on a contrast gain control model of the HVS. He perform a conversion from the Y'Cb'Cr' space color defined in the ITU-R Recommendation 601, to an opponent color space (B-W: Black-White, R-G: Red-Green and B-Y: Blue-Yellow) based on the HVS cones sensibility to each color component. They take into account in their color space transformations the behavior of conventional CRT (Cathode Ray Tube) displays.

In [76, 74] authors convert the reference and the distorted image into the YOZ color space, where Y is the luminance expressed in $candela/m^2$, O is an opponent color channel calculated with a specific conversion matrix, and the Z channel is the blue channel given by the CIE Z coordinate. This transform also includes gamma transformation and a linear color transform.

Nevertheless some other authors do not perform any color conversions or transformation, they in fact retain only the luminance information in order to reduce the computational cost of their proposed metrics. Authors in [1] introduce a Perceptual Blocking Distortion Metric based in the model proposed in [72]. They perform also the most important steps from the ESF, as frequency decomposition, contrast sensitivity filtering, contrast gain control, error detection and pooling. Regarding the color conversions authors argue that only if the metric precision is a critical issue then a color conversion as in [79] is worthwhile, as it has been shown [80] that it is possible for the vision model to work on the luminance (Y) component only, without a dramatic degradation in prediction accuracy. They propose also that the contrast sensitivity band-pass filtering can be applied only to the luminance channel, based on the fact that color contrast sensitivity is rather low for higher frequencies, reducing therefore computational costs.

Another type of pre-processing step is the need to convert the digital pixels (stored in the computer memory) into luminance values of pixels on the display device, through point-wise non linear transforms. Different gray-level transformations or corrections are applied as pre-processing step in order to account to contrast adaptation to luminance conditions.

Finally, the reference and the distorted images or videos need to be converted

into corresponding contrast stimuli to simulate light adaptation. There is no universally accepted definition of contrast for natural scenes. Many models work with band-limited contrast for complex natural scenes [81], which is tied with the channel decomposition. In this case, the contrast calculation is implemented later during or after the channel decomposition process.

CSF

The CSF can be implemented in the channel decomposition step by the use of linear filters that approximate the frequency responses to the CSF. Like in [82] that is based in a local contrast definition and where a spatio-temporal three dimensional filter bank is applied to the image, decomposing it in different frequency perceptually channels. The filter bank design takes into account subjective psycophysical experiments in order to fix the contrast sensitivity for each frequency range and orientation, an so, the frequency channel decomposition includes the contrast sensitivity function.

But most of the metrics choose to implement the CSF as weighting factors that are applied to the channels after the channel decomposition, providing for each channel a different perceptual sensitivity. In chapter **??** we will discuss how to introduce the CSF after the decomposition step but in the image encoding scope.

Decomposition

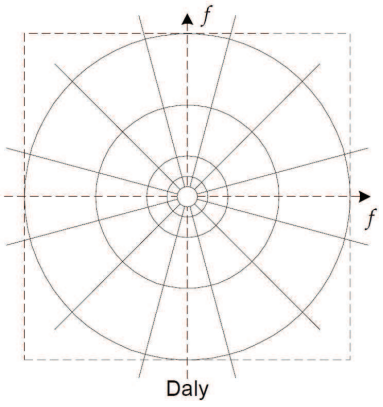


Figure 1.23: Daly frequency decomposition model

Transformations from the image spatial domain into the frequency domain has been extensively use in the literature in image and video coding algorithms. The

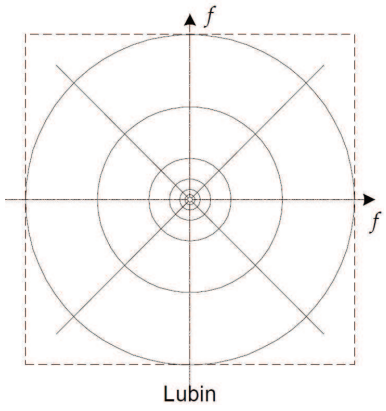


Figure 1.24: Lubin frequency decomposition model

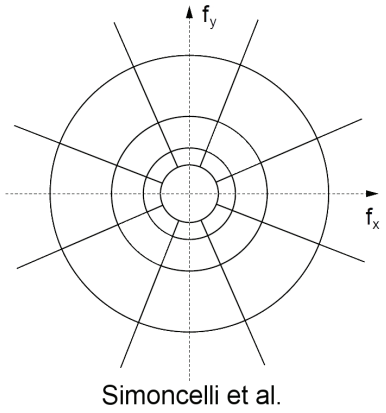


Figure 1.25: Simoncelli et al. frequency decomposition model, Steerable Pyramid

most widely used frequency transforms are the Discrete Cosine Transform (DCT) and the Wavelet transform. This simple transforms have been reported due to their suitability for the codification process and certain applications, rather than their accuracy in modeling the cortical neurons; their models are not close enough to the channel decomposition that our HVS does while process the incoming signal from our eyes. Nevertheless some metrics use DCT [75] or Wavelt [63] frequency decomposition with good correlation with MOS values.

Quality metrics, that try to emulate, as accurate as possible, the way in that our HVS assess the quality of the viewed scene, use more complex models of this HVS frequency channel decomposition, but taking into account the constraints of application and computation. Depending also on the metric type and the type of distortions it handles, metrics use different different channel decomposition

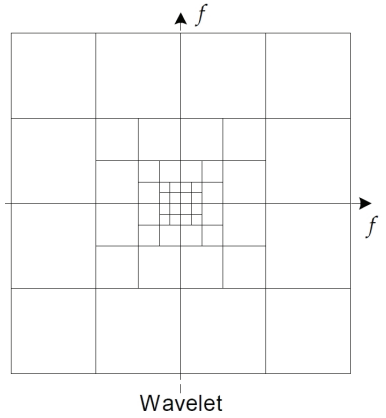


Figure 1.26: Wavelet frequency decomposition model

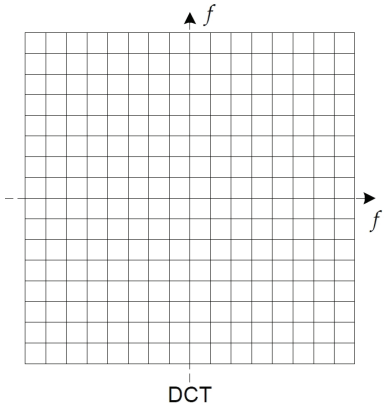


Figure 1.27: DCT frequency decomposition model

models.

Cortical receptive fields are normally represented by 2D Gabor functions, but the Gabor decomposition is difficult to compute and is not suitable for a good computational light implementation and for some operations as invertibility, reconstruction by addition, etc.

Normally, frequency decomposition is produced by a filter banks in which design must be incorporated spatial location, spatial frequency and orientation in order to resemble the HVS frequency and orientation channels. This filter bank design differs among authors. From a practical and implementation point of view several authors have implemented pyramidal filter structures. In [83] Watson modeled a frequency and orientation decomposition with similar profiles than the

2D Gabor functions but computationally more efficient. Other authors like Lubin [37], Daly [84], Teo and Heeger [72] and Simoncelli et al. [85] provided different models trying to approximate as close as possible to the HVS channel decomposition avoiding prohibitive implementation issues. In [85] Simoncelly proposed the steerable pyramid which is a frequency multi-scale and multi-orientation image decomposition that is invariant to translations and rotations of the stimuli, without aliasing effect and invertible. In figures 1.23 to 1.27 some of this channel decomposition models are shown.

There are also some models that cover temporal frequencies decompositions in order to account for the characteristics of the temporal mechanisms in the HVS [79, 82]. The design of temporal filter banks is normally implemented using Infinite Impulse Response filtesr (IIR) that give a delay only of a few frames, other authors use Finite Response Filters that although having a bigger delay are simpler to implement.

Although the use of that sophisticated channel decomposition models is commonly used in QAMs, normally simpler transforms like DCT or Wavelet are still employed in the design of image or video codecs due mainly to its reduced computational cost.

Error Normalization and Masking

As explained in 1.3.8 masking occurs when a stimulus that is visible by itself cannot be detected due to the presence of another stimulus. Sometime facilitation occurs, that is when a non visible stimulus becomes visible due to the presence of another.

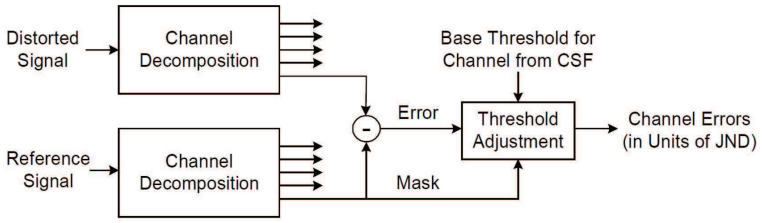


Figure 1.28: Typical implementation of masking in quality metrics

Most of the HVS models in this framework, implement error normalization and masking in the form of a gain-control mechanism, using the contrast visibility thresholds in order to weights the error signal for each channel, see figure 1.28. Some metrics [73], normally due to complexity and performance reasons, use only intra-channel masking, i.e. masking occurs only in each region

of the decomposed (frequency and orientation) spectral domain, while other models [72] include inter-channel masking as there are evidences that channels are not totally independent in the HVS.

The visibility threshold adjustment at a point is calculated based on the energy of the reference signal (or both the reference and the distorted signals) in the neighborhood of that point, as well as the HVS sensitivity for that channel in the absence of masking effects (also known as the base-sensitivity). For every channel the base error threshold (the minimum visible contrast of the error) is elevated to account for the presence of the masking signal, and for this masking elevation several masking models are typically used. The elevated visibility threshold is then used to normalize the error signal. This normalization typically converts the error into units of Just Noticeable Difference (JND), where a JND of 1.0 denotes that the distortion at that point in that channel is just at the threshold of visibility.

Some authors [86] include also in this stage the luminance masking also called light adaptation. Detection threshold for a luminance pattern depends upon the mean luminance of the local image region. So, the brighter the background is the higher the luminance threshold is. Up to a variation of 0.5 log units in the luminance threshold might be expected to occur within an image due to the mean luminance of the block for which it is calculated (assuming a block basis image encoder). Watson propose a power function for approximate the luminance threshold for a DCT block. In [76] a local contrast calculation is included for every DCT block converting each DCT coefficient in a value in the range from -1 to 1, that expresses the amplitude of the corresponding basis function to the average luminance in that block.

In [33, 87] we can find comparisons of different masking models and some considerations about how to include them into an image encoder. In [88] authors propose a contrast gain-control model of the HVS that incorporates also a contrast sensitivity function for multiple oriented bandpass channels.

Error Pooling

The last step in the process is the error pooling which is the process of combining the error signals in different channels into a single distortion/quality interpretation giving different importance to errors depending on the channels. For most quality assessment methods, a Lp norm or Minkowski norm is used for error pooling expressed as in equation 1.3. Where $e_{l,k}$ is the normalized error of coefficient k at frequency level l and β is a constant value lying between 1 and 4. Importance weights can also be given based on the visual importance of different regions in the image.

$$E\left(\{e_{l,k}\}\right) = \left(\sum_{l}\sum_{k}\left|e_{l,k}\right|^{\beta}\right)^{\beta} \tag{1.3}$$

Most of the previously cited metrics are FR metrics and follow the functional stages of the Error Sensitivity Framework although with variations. This schema, specifically the summation or pooling stage, allow the metrics to produce spatial error maps, frame-level distortion scores and sequence-level distortion scores. In these sense an image quality assessment metric can be use directly to rank video sequences. For the time domain some metrics use temporal HVS models or information to accurately reproduce human scores while others simply provide their sequence quality value as a frame-quality average.

Now, we will summarize the most relevant and cited metrics of this framework.

In [72] model, Teo and Heeger include basically all steps from EFS and is one of the first reference metric of this framework. Its model is based in the analysis of the responses of single neurons in the visual cortex of the cat, where a contrast gain control mechanism keeps neural responses within the permissible dynamic range while at the same time retains global pattern information. They perform a Quadrature Mirror Filter (QMF) frequency decomposition. The gain control mechanism is realized by an excitatory nonlinearity that is divided by a pool of responses from other neurons. The distortion measure is then computed from the resulting normalized responses by a simple squared-error norm as explained before.

The Moving Picture Quality Metric (MPQM) [82, 73] is a FR metric that pre-process the sequences in blocks, doing a coarse segmentation of regions, uniform, pattern and borders, in order to fix the base masking threshold for each image block. Frequency decomposition is based on a local contrast definition and Gabor-related filters for the spatial decomposition, it uses an isotropic filter for low frequencies regardless the orientation and for the frequency bands of 2,4,8 and 16 cpd and another filter for each orientation (0, $\pi/4$, $\pi/2$ and $3\pi/4$). The 17 filtered spatial decomposition is followed also by two temporal mechanisms, as well as a spatio-temporal CSF and a simple intra-channel model of contrast masking. The masking mechanisms consist of dividing the filtered error signal (original filtered minus distorted filtered) by the detection threshold getting this way data in "'units above threshold"'. Data from each channel is gathered together in a pooling step. They provide results for a global metric and for more detailed metrics for each of the basic image components: uniform areas, contours and textures. The global metric takes also into account the focus of

attention computing the sequence in three-dimensional blocks accounting for persistence of the images on the retina. Pooling this three-dimensional blocks the global distortion measure is given. The final distortion measures (global and components) can be obtained in "'Visual Decibels"', expresed in the commonly used decibels (dBs) or in a quality rating on a 1 to 5 scale resembling the MOS scale.

Based on self developed non-linear and supra-threshold contrast perception model authors in [75] propose the use of a FR metric, working in the DCT domain, that deals with a wider range of distortions than other model based metrics. Their model is based on experimental perception results, so it models as a whole the HVS, including the effects from photoreceptors to the post-transform suprathreshold non-linearities. They argue that such a model works better than models that are base on a stage-after-stage sequential model based on disconnected characteristics of the HVS. Based on the fact that the HVS maps continuous contrast range into a finite set of discrete perceptions, they model the bit allocation properties of the HVS as a redundancy removal process analogous to vector quantization. Their experimentally parametrized Information Allocation Function (IAF) model, is based on the idea that if the HVS allocated more information in one area (frequency and orientation), more visual importance is then given to that area. Their IAF value, that includes not only sub-threshold or at threshold behavior of HVS but also the reactions to supra-threshold impairments, is used to weight the DCT coefficients, and by measuring the differences between the perceived images (original and distorted are processed with the IAF) a subjective difference between both image is given.

Following the ESF framework stages, in [86] Watson introduced the DCTune metric a FR metric for monochrome images, tested with the JPEG image compression standard, which was extended in [89] for color images and in [76] for color video sequences with the name of Digital Video Qualtiy (DVQ). The method treats each DCT coefficients as an approximation to the local response of a visual channel. For a given DCT quantization matrix the DCT quantization errors are adjusted through each one of the ESF stages (contrast sensitivity, light adaptation and contrast masking) and pooled non-linearly over the blocks of the image. This process results in a 8x8 "'perceptual error matrix"' which is further pooled again for each block to give the final total perceptual error. In [86] author argue in favor of an image dependent quantization matrix giving arguments against an image independent quantization matrix. He propose a method, that following each of the ESF stages, obtains an visually optimum (at threshold) quantization matrix for a specific image and bitrate. In [76] author include the results of measurements of visual thresholds for temporally varying samples of DCT quantization noise in order to extend its previous metric to the time domain.

In [74] authors extended the previous work providing also results from subjective tests.

Although not following all the stages of the ESF, authors in [8] propose a FR measuring tool for MPEG-2 video sequences. Their proposal is different as they include a "Cognitive Emulator" stage after the "Distortion Weighting" stage. This cognitive modeling of quality assessment is seldom included in quality assessment metrics, and therefore this proposal is interesting because not only include a low-level model of HVS but also try to model high-level cognitive decision stages.

In the Distortion Weighting stage, authors apply a low-pass filter to the original and distorted sequence, with similar response as the CSF. Then, with the aid of a edge detection step, that runs on the original image, a simplified masking model is applied. The masking is applied in the space domain by modifying the luminance values of the neighborhood (± 5 pixels) of the edges, being maximum at the sharp luminance transition. The masking function is applied for vertical and horizontal edges and is composed as a combination of local masking functions for the pixels in the aforementioned neighborhood. Prior to the Cognitive Emulator authors obtain what they call the IPQ (Instrumental Picture Quality). IPQ is a normalization and mapping of the PSNR to the visual rating. As subjective rating of quality saturates above and below certain quality values, they simply apply this saturation effect to the calculated PSNR of the distorted image, getting this way their IPQ. Their saturation limits were fixed at 20 and 50 dB. The Cognitive stage is a predictor of the subjective results from SSCQE subjective evaluation tests on video sequences. Authors propose a model to reproduce the decision making tasks involved in a SSCQE test. Their Cognitive model try to mathematically include the biased judgment that could be expected as result of the rapid picture quality variation in the video sequence and the need to rapidly decide the perceived quality. Based on the short-term human memory behavior, the influence of strong stimulus that appears in a frame, persists during several frames. When another strong stimulus occur within an interval shorter than the memory interval this two stimuli may merge and normally mask the quality of frames inside the two distorted frames. Due to the presence of the distorted frames, the quality of the frames inside is judged to be worst than it would be in the absence of the distorted frames. This fact is modeled by the authors as a smoothing stage that modify the IPQ value of frames between frames with lower IPQ value. The perceptual saturation is also included in their model by normalizing the IPQ values in the range of 0.0 to 1.0. After the Smoothing and the Perceptual Saturation stages an Asymmetric tracking stage is performed. This stage takes into account the fact that observers respond decisively and quick to degradation in picture quality, but hesitate and slow in the case of picture improvement. They model the subjective gain and losses response

by modifying asymmetrically the value of the IPQ values to account for this fact. The final stage is to delay in time the point where the modified quality value is applied in the sequence due to the human response time that was previously estimated (averaged) as 1 second. All these Cognitive stages try to synchronize the video distortion with the SSCQE data.

Author in [79] propose the Perceptual Distortion Metric (PDM) a FR metric for color video sequences, based on a contrast gain control model of the HVS. He perform a conversion from the Y'Cb'Cr' space color to an opponent color space as pre-processing stage. This metric propose a separated temporal and spatial frequency decomposition. In the research of the temporal mechanisms in HVS there is a consensus of the existence of at least two filtering stages, a low-pass and a band-pass referred as sustained and transient channels. Winkler uses two IIR filters to model these stages applying the low-pass filter to all three color channels while the band-pass filter is applied only to the luminance channel to reduce complexity. The spatial decomposition is implemented with the steerable pyramid transform proposed by [85] which has the advantage of being rotation-invariant, self-invertible and because minimizes the amount of aliasing in subbands, but requires higher computational load. CSF is implemented as a weighting process after subband decomposition. Masking is implemented as an extension of Watson [86] masking model to color images and to video sequences. In [80] the author tested the PDM metric with different color models. Using the CIE L*a*b* and CIE L*u*v* models with the metric has better correlation with human scores. He concluded also that using a luminance only model produced slightly lower correlations but the slight increases in accuracy of the color versions may not justify the double computational load imposed by the full-color PDM.

Encoding images giving more bits (information) to the correct Regions Of Interest (ROI) and discarding less important information from peripheral regions can be perceptually improved by maximizing quality value given by a foveated quality metric. Therefore, some metrics use models of the HVS that include foveation (see 1.3.2) in their design. In [77, 39] the Foveated Wavelet Image Quality Index (FWQI) is presented. FWQI is a FR metric working in the wavelet domain and based on the fact that the HVS is highly non-uniform in sampling, coding and processing. The HVS spatial resolution is higher around the fovea and decreases rapidly with increasing eccentricity. The reason of using a wavelet decomposition for this metric is because wavelet analysis delivers a convenient way to simultaneously examining frequency and spatial information. The design of this metric include information about the space variance of the CSF, spatial variance of the cutoff frequency and information about the variation of the human visual sensitivity in different wavelet subbands. The distance to the image

and the display resolution plays also an important role. The perceptual importance of each wavelet subband is taken from the model in [28], which fixed the error sensitivity for each subband based on experimental results. Authors combine this model with a model of the distance of each wavelet coefficient to the foveation point in spatial domain, obtaining after pooling their FWQI.

In [1] authors propose a blocking impairment metric, the Objective Blocking Rating (OBR) and the Perceptual Blocking Distortion Metric (PBDM) based in the OBR. PDMB is a FR metric based in the [72] HVS model with the modifications made in [82], that include temporal filters and CSF, and also with the color extensions made by [79]. This extended model was finally modified to change the gain control stage to the one proposed by [88]. All the stages in the model clearly explained and slightly simplified to reduce computational effort. After some parametrization authors get the same correlation with MOS values than the PDM metric, but with lower computational cost.

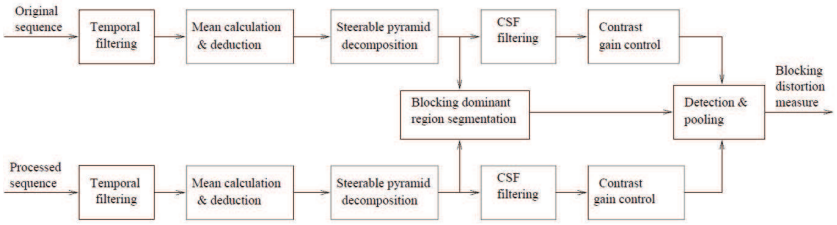


Figure 1.29: Block diagram of the PBDM [1]

The main steps of [1] can be shown in 1.29. The Steerable Pyramid is used to perform the frequency decomposition, but only to a central region of the image in order to avoid boundary effects. The CSF is then implemented as a weight factor that multiplies each subband in the wavelet domain. The CSF weighted coefficients are then passed to the gain control mechanisms that squares and normalize the coefficients. As known, the *LL* subband holds the low-pass band. It is important to notice that authors pre-process the frames in order to be able to pass the gain-control stage to this subband, by substracting the mean value is subtracted to each pixel in the frame (in the spatial domain) before the frequency decomposition. This pre-processing step is needed therefore, in order to prevent the accumulation of energy into the low-pass band, which could produce that the magnitude of that coefficients fall out of the the dynamic range of the gain-control stage. A final pooling stage simulates the integration process of the HVS obtaining finally the perceptual distortion map, with the same size as the original frame, assign to each pixel the perceptual distortion at that spatial location.

As exposed in Fig 1.29 authors propose and introduce as an additional blocking stage, so that their algorithm produces a blocking region map. They also provide a method to calculate the ringing artifacts produced after the frequency decomposition, but as ringing is produced due to edges reconstruction errors should not be considered as blocking artifacts, so that ringing areas are excluded from the blocking region map. Both algorithms rely on experimentally adjusted thresholds. Authors averaged the summed blocking distortion by the number of frames and experimentally adjust the dynamic range of the metric in a scale of 1 to 5. Blocking distortion is calculated in the previously segmented "'blocking dominant region"'.

As the focus of attention of viewers is located mainly in faces and moving objects authors in [78], although not proposing a novel metric, they combine the use of two quality assessment metrics in order to achieve the global quality rating of a video sequences. When focus of attention is located on a particular area of the scene the background or the rest of objects in the scene are coarsely processed. They combine the previously commented FR PDM metric, which is based on a HVS model and NR [90] metric to measure the influence of blockiness, blur and jerkiness artifacts. The combined metric is guided from a semantic segmentation of images. The semantic segmentation is produced mainly for people faces. When focus of attention is placed in moving objects, then background objects or those with different velocities are processed less accurate also. In [48] a spatio-temporal CSF model that account for this is presented.

Authors in [63] propose an interesting proposal of two metrics, a FR and a RR one for video sequences, being based both metrics, on the same HVS model. Their model follow all the aforementioned stages as, color space conversion, temporal filtering, spatial filtering, contrast computation masking and summation. As they point out, the use of a RR or a NR metric that is specifically designed for catch some impairments, as blocking or blurring, have the disadvantage of not being able to determine if one potential artifact is part of the sequence or the result of the compression process of a new generation of codecs or algorithms. Therefore they based their RR in the same HVS model than the FR one, but working with a reduced bandwidth version of the reference sequence. This reduction can be scaled up to FR, adapting to the available bandwidth. Although their model is based on previous HVS models, the parametrization that authors perform to the model is guided by the responses to natural video frames rather than by the responses to simple visual stimuli such as sinusoidal gratings. In addition authors propose a method to perform a perceptually driven rate control based on a previous work [91] and using the new RR metric as distortion measure in the rate control algorithm.

HVS Properties Framework

In this framework we include other types of metrics, that although are not based on a specific HVS model, are still inspired in the HVS in the sense that their design takes into account some of the aforementioned HVS properties. We also include here, those metrics that are build to detect specific impairments produced by any of the processing stages of images and videos, like quantization, encoding, transmission etc, by analyzing different image properties.

The Institute for Telecommunication Sciences (ITS) presented in [92] an objective video quality assessment system that was based on human perception. Instead of following stage by stage one of the HVS models, they extract several features from the original and degraded video sequences. That features were forward statistically analyzed in comparison with the corresponding human rating extracted form subjective tests. This analysis provide the parameters that adjust the objective measures for these features and after being combined in a simple linear model, they provide the final predicted scores. Some of the extracted features require the presence of the original sequence while others are extracted in a no reference mode. The proposed metric exploits spatial and temporal information. The processing include Soebel filtering, Laplace filtering, fast Fourier transforms, first-order differencing, color distortion measures and moment calculation.

Based on previous works, the ITS in [62], proposed a RR metric for in-service quality monitoring system. Their metric is build on a set of spatio-temporal distortion metrics that can be use for monitoring in-service of any digital video system. Authors expose that a digital video quality metric, in order to be widely applicable must accurately emulate subjective responses, must work over the full range of quality (from very low bit rate to very high), must be computationally efficient and should work for end-to-end in-service quality monitoring. The metrics presented in their work are based in extracted features from the video sequence as in [92], and in order to satisfy the last condition (to be able to work in in-service monitoring systems), these features, extracted from spatio-temporal regions, are sent, compressed following the ITU-R Recomendation BT.601, through an ancillary data channel so that it can be continuously transmitted. In the paper the authors describe these spatio-temporal distortion metrics in detail so that can be implemented by researchers.

Later, through The National Telecommunications and Information Administration (NTIA), the same authors, proposed the General Model of the Video Quality Measurements Techniques (known as VQM metric) for estimating video quality and its associated calibration techniques. This metric was

submitted to be independently evaluated on MPEG-2 and H.263 video systems by the Video Quality Experts Group (VQEG) in their Phase II Full Reference Television (FR-TV) test. The VQM, that is based on the same algorithms used in their previous works [92, 62] was standardized by the VQEG and a technical report [93] was supplied with a full description of the metric and all its operation modes. This metric was later summarized in [94]. As mentioned before the VQM uses RR parameters that are extracted from optimally-sized spatio-temporal regions of the video sequence. The ancillary channel and the calibration techniques require at least a 14% of the uncompressed sequence bandwidth. Information is sent through that channel. Although being conceptually a RR metric was submitted to the VQEG FR-TV test because the ancillary channel can be use to receive more detailed and complete references from the original frames, even the original frames themselves. The VQM with its associated calibration techniques comprise a complete automated objective video quality measurement system. The calibration techniques include spatial alignment, valid region estimation, gain and level offset calculation and temporal alignment. Finally in [95] authors reduce the requirements of some of the features extracted in the NTIA General Model in order to achieve a monitoring systems that uses less than 10 kbits/s of reference information.

In [96] authors propose a NR metric for blocking artifacts in images. Previous NR blocking metrics measured the amount of blocking by using a weighted mean-squared difference along block boundaries [97]. This method can produce situations in that even the original image can be evaluated as blocky. Authors propose to treat the distorted image as a pure non-blocking image that is interfered with a pure blocky signal, and the key of the metric is to measure the power of that blocky signal. They define an ideal 1-D blocky signal that is suppose to be interfering the original image for each row and column. For measuring the amount of blocking they use a power spectrum estimator of the image in the Fourier domain, i.e. after applying the FFT. A final weighting and summing stage, that processes row and column information, produces the final blocking measure.

Authors in [98] propose another NR metric for blocking artifacts, this work was extended in [99]. Their metrics works in the DCT domain. They first define a 2-D step function for modeling an overlapping block that is made off the bottom and upper part of vertically adjacent blocks, or left/right for horizontal adjacent blocks. Once they have modeled the 2-D step function of that “overlaping block” and are able to measure the amount of edge activity (blocking) in the DCT domain, they include the luminance masking and the texture masking in the process. Although more accurate models have been proposed in the literature, they propose a simple model of texture masking

artifacts to facilitate real-time operations, using the amplitude of the 2-D step function and the amount of blocking measured for the horizontal and vertical edge activity. For luminance masking they adopt the model proposed in [100]. Finally they produce a map of "artifact visibility" for the whole image, so that block artifacts reducing algorithms can adaptively work according to local visibility. They also provide a combined numerical value as a global blocking artifacts measure in the image.

A NR metric for blocking and blurring and specifically designed for JPEG compressed images is presented in [101] with low computational cost. Authors provide a Matlab implementation of the metric and the value for their model parameters obtained so that the results can be reproduced. The metric measures blocking and blurring combining both together to get the final image score. First they calculate for each row a new row that holds the differences with the previous row. This "'differences image"' is used to calculate next values. The blockiness measure is estimated as the average differences across block boundaries and the blurring is calculated using the activity of the image signal. The activity is calculated using the average for in-block differences and the zero-crossing rate for each block. A zero-crossing occurs when for a "differences row" the difference value for a specific column crosses zero, i.e. previous column has positive value and next column negative or vice versa. Finally the blockiness and the two activity measures are modeled in an equation whose parameters are obtained by fitting the MOS values of various test image sets.

A NR perceptual blur metric is presented in [102] that is based on the analysis of the spread of the edges in an image. They argue, based on a correlation with MOS values, that measuring the spread of vertical edge is sufficient to model the perception of blur, avoiding to repeat the measures for horizontal edges or in the direction of the gradient of that edges. They use a Sobel filter to detect vertical edges and measure the local blur for each row as the width of the edge. Averaging this local blur for all the edge locations on the whole image they get the final blur measure. To detect the width of each edge detected with the Sobel filter, the beginning and end pixels are determined by searching around the edge location the local maximums and minimums for each row. Their proposal has low computational complexity and its performance is independent of the image content.

In [103] the same authors extended their work to include the aforementioned NR blur metric with a FR Blur metric and a FR Ringing metric. The proposed metrics are defined in the spatial domain with a very low complexity and are based on the analysis of the edges in an image. The blur metrics measure the spread of the edges and the ringing metric measures oscillations around edges. In the FR

version the edge used for their algorithm are those from the original image while in the NR version the edges are obtained directly form the processed or compressed image. The ringing metric is based on the FR blur metric. From the wavelet decomposition filters they obtain a fixed ring-width. The edge width, from blur metric, is substracted to that ring-width, this width which is the distance around the edge (left and right) where differences (oscillations) with original image are locally measured for each edge position. The difference between the maximum and minimum difference in the ring-width (left and right) is multiplied by the ring-width itself giving the amount of ringing for each edge position. Averaging for all edge positions in the image they obtain a global ringing measure. They finally combine both metrics (blur and ringing) to a FR quality metric.

The Reduced Reference metric called HIQM (Hybrid Image Qualitiy Metric) is proposed in [104] is a weighted sum of different image artifact measures (smoothness, blocking, ringing, masking and lost block/pixel). The blocking measurement is based on the algorithm proposed by Wang et. al. [96, 101]. The blur measurement algorithm is based on previous work in [102]. They use the metric proposed in [105] to detect ringing and lost blocks by measuring the edge activity and the gradient activity that is higher in the distorted image due to the apparition of false edges. Finally masking detection is based on the global contrast measure of the image that is in turn based on the standard deviation of the first-order image histogram that is used to measure the average brightness of the image. A weight is given for each distortion and an averaged weighted sum produce the final quality value of the metric. The weights are empirically obtained in order to achieve a good correlation with PSNR.

The proposal of [106] include another way to assess the quality of images. In this case, images to be judged are improved versions of the original ones, i.e. they try to predict the quality of enhanced images. Authors argue that the Error Sensitivity approach or the use of RR or NR metrics that are based in properties of the distorted image are not suitable for this task because those methods are designed to assess quality of degraded images. So they propose to use a neural network that has been trained to predict the final quality of the enhanced images as it would be judge by human assessors. The inputs to the neural network are numerical values corresponding to several objective properties of the enhanced image. These values are determined at the signal level, i.e. are based on pixel values that are extracted block by block (block size 32x32 pixels). These features describe the image content in terms of luminance distribution, spatial orientation, frequency energy distribution, etc.

As in other proposals, authors in [107] propose the use of a RR metric to assess the quality of a video sequence. They use image properties or indicators to

measure differences between the original and distorted image that are encoded and transmitted with the video sequence. So at the decoder side the same properties are obtained from the distorted image and compared with the original ones. Authors use this RR metric in combination with another NR metric to assess quality of video streaming over IP networks. The RR metric accounts for image quality while the NR metric accounts for transmission quality. The basic indicators for the RR metric include the Estimated Additive Gaussian Noise power level (based on Wiener filtering), the Impulsive Noise power level estimation (based on median filtering), Blocking and Blurring artifacts (based on [96, 101]) and finally statistics of Ringing Artifacts (based on a Perona-Malik filter). These properties are embedded in the coded bitstream. The NR component mainly refers to the impact of temporal resolution reduction, packet losses, latency and delay jitter. Although packet loss and out of sequence ratios can be derived by gathering the communication channel output, authors use only the decoded information to detect these effects.

Finally other metrics that take advantage of the the contrast masking effect of the HVS are included in this framework. So, we can find metrics based on watermarking techniques that analyze the quality degradation of the embedded image [108]. Also, in the metric presented in [109] based in a new concept named "Quality-aware image", authors extract some features of the original image that are embedded into the image as invisible hidden messages. When the distorted image is received the loss of parts of that hidden features yields to a objective measure of the quality of the received image.

Statistics of Natural Images Framework

Some drawbacks of the Model Based HVS framework are reviewed in [2, 110]. Some of these drawbacks are, for example, that the HVS models work appropriately for simple spatial patterns, like pure sine waves, however when working with natural images, where several patterns coincide in the same image area, then their performance degrades significantly. Another drawback is related to the Minkowsky error pooling, as it is not a good choice for image quality measurement. As authors show, different error patterns can lead to the same final Minkowsky error. Also the HVS Model based framework is designed to estimate the threshold at which a stimulus is just barely visible. These subjectively measured threshold values are then used to define error sensitivity measures as the CSF and various masking effects. But most of the impairments produced while processing images are above this thresholds, i.e. are clearly visible, so it is not clear that the near-threshold models can accurately assess suprathreshold distortions. Some studies try to include suprathreshold psychophysics for

analyzing image distortions [111, 112, 113].

Therefore, several authors argue that the approach to the problem of perceptual quality measurement must be a top-down approach, analyzing the HVS to emulate it at a higher abstraction level. The authors supporting this approach, propose to use the statistics of the natural images. In [114] a review of recent Natural Scenes Statistics (NSS) models is presented.

Some of them propose the use of image statistics to define the structural information of an image. When this structural information is degraded, then the perceptual quality is also degraded. In that sense, a measurement of the structural distortion should be a good approximation to the perceived image distortion. These metrics are able to distinguish distortions that change the image structure from distortion that do not change it, like changes in luminance and contrast.

In [2, 115] authors define a Universal Quality Index that is able to determine the structural information of the scene. This index models any distortion as a combination of three different factors: a) the loss of correlation between the original signal and the distorted one, b) the mean distortion that measures how close the mean of the original and distorted version are, and c) the variance distortion that measures how similar the variances of the signals are. The dynamic range of the Quality Index i [-1,1] being 1 the best value, when the signals are identical. They apply this index in a 8x8 window for an image obtaining a quality map of the image. The overall index is the average of the quality map.

Authors in [110] further improve their previous quality index proposing the SSIM (Structural SIMilarity) quatliy index. This metrics, based in the Universal Quality Index [2, 115] works in the spatial domain. They expose that the index get better results if it is applied locally and then averaged rather to apply it over the whole image. Applying the SSIM locally reduces the foveation effect, as at typical viewing distances only a part of the image is perceived with high resolution, and can provide a spatially varying quality map of the image. Instead of applying it in a 8x8 block basis as in their previous work, which produces blocking effect, they use a 11x11 circular-symmetric Gaussian weighting function. They use the Mean SSIM (MSSIM) index to evaluate the overall image quality. Due to the existence of the quality map, the quality of Regions Of Interest (ROI) can be easily computed by averaging the quality in that regions. Also several weighting functions can be applied to the local quality index in order to adapt to any application, however they use a uniform weighting. This work was later fully explained as a book chapter in [116].

Authors in [117, 118] proposed a video quality metric following a frame by

frame basis. Authors apply the SSIM index locally in 8x8 blocks randomly selected to reduce computational costs. They apply the SSIM index to the Y,Cb and Cr color components independently and obtaining the global color SSIM index using a weighted summation. Using statistical features like mean and variance they classify the blocks as smooth region, edge region or texture region. Results of all the selected areas are averaged to give the frame quality value. This value is further adjusted based on the overall blockiness of the image and the motion factor. The blockiness and blurring are evaluated globally for each frame using the NR metric proposed in [96]. Instead of using a uniform weighting factor while averaging the randomly selected blocks, they assign different weights based in the local luminance, for example, as dark areas attract hardly the attention of the viewer these areas get a lower weight. Authors also perform a second adjustment based on how the blur distortion is considered depending on the motion in the scene. The motion information is obtained by a simple block-based motion estimation algorithm with full pixel resolution. The final video sequence quality index is the average of the frames quality values. In a still or low motion frame, severe blurring artifacts are very annoying, but in a large motion frame the same amount of blur is perceived as less important because motion blur occurs at the same time. They give different weights according to the type of the frame motion.

In [119] extended their SSIM to a new Multi-Scale Structural SIMilarity (MS-SSIM) model. The new proposed multi-scale analysis runs a low-pass filter to the images (original and distorted versions) and a downsampling process to the filtered images iteratively. Then at each of the resulting scales the SSIM index is applied. After M-1 iterations the Scale M is obtained being the original resolution the Scale 1. At each scale the contrast comparison and the structure comparison of the SSIM is applied whereas the luminance comparison is applied only at Scale M. The final multi-scale SSIM index is obtained by a weighted combination of the comparison operators. Different weights can be applied a to each scale, in the same sense as the CSF apply different weights to each frequency subband, they uniformly weight each scale. They perform a subjective test in order to detect the perceptual importance distortions (in increasing grade) applied at each scale. The results of this subjective test provided the perceptually adjusted weights for each scale. The reason why authors did not use the CSF for this task, is because it is typically measured at visibility thresholds levels and using only simplified stimuli (sinusoids) and the purpose of the new MS-SSIM is to compare the quality of complex structured images with distortions above threshold.

As stated in [120] the main drawback of the spatial domain SSIM algorithm is that it is highly sensitive to translation, scaling and rotation of the image. So, in this work [120] authors presented the Complex Wavelet SSIM (CW-SSIM) which

extend the SSIM method to the complex wavelet transform domain and make it insensitive to non-structural distortions like zoom, rotations and translations produced by movements of the acquisition devices. This insensitivity works only if this movements or zooms are smaller than the used wavelet filters.

In [121] authors propose a general adaptive linear system framework that is able to decompose the distortion between two images into linear combinations of the constituent distortions. One linear combination corresponds to non-structural distortions like luminance and contrast changes, gamma distortions and horizontal and vertical translations. It is obtained in a pre-procesing step where the weights for each type of distortion is also computed. The other combination corresponds to structural distortions. A frequency decomposition method, based on the DCT transform matrix, is applied to obtain the structural distortions. With the weighted combination of the two types of combination a QAM is proposed.

Other authors use also statistics of the scene in a different way. They state that the statistical patterns of natural scenes have modulated the biological system, adapting the different HVS processing layers to these statistics. First a general model of the natural images statistics is proposed. The modeled statistics are those captured with high quality devices working in the visual spectrum (natural scenes). So, text images, computer generated graphics, animations, draws, random noise or image and videos captured with non visual stimuli devices like Radar, Sonar, X-Ray, etc. are out of the scope of this approach. Then, for a specific image, the perceptual quality is measured taking into account how far its own statistics are from the modeled ones.

In [122] a statistical model of a wavelet coefficient decomposition is proposed, later in [123] a RR Image Quality Assesment metric (RRIQA) is presented. Authors use a model of the statistics of natural images in the wavelet transform domain. They work with the steerable pyramid wavelet transform from [85] and use the Kullback-Leiber Distance (KLD) to measure how different are the marginal probability distributions of wavelet coefficients in the reference image and distorted images. This is used as measure of distortion. They find that several well known types of image distortions produce significant changes in the wavelet coefficient histograms what is detected by the metric. They do not assume any distortion model, so the proposed method is potentially useful for for a wide range of distortion types. The marginal probability distribution from the distorted image is obtained directly from the decoded wavelet coefficients, but the marginal distribution from the reference must be transmitted to the receiver as RR data. If the histogram bin size is small then the bandwidth required to transmit the RR features is very demanding, but if the histogram bin size is large then the accuracy of the KLD is reduced. But they send only three parameters as

RR data. The cue is that the marginal distribution of the coefficient in an individual wavelet subband can be modeled as a two-parameter Generalized Gaussian Density model (GGD) as they refer. The third parameter is the prediction error between the original distribution and the GGD distribution. So, in the receiver side using the GGD parameters and the error prediction the marginal distribution of the reference image can be reconstructed. These parameters are computed and sent for each wavelet frequency subband.

In [124] authors propose an NR metric (NRJPEG200) that uses a statistics of natural images model in the wavelet domain [125, 126] in conjunction with information of the distortion model of the JPEG2000 encoder. With both information they build a simplified model that characterize images compressed by JPEG2000 as well as uncompressed natural images. The statistical model predicts the wavelet coefficient's magnitude conditioned on a linear prediction of the coefficient. The linear prediction is calculated based in two image dependent estimated thresholds and the relationship of the coefficient with its parent, grandparent and its neighbors. The quantization of wavelet coefficients produces a reduction of the significant coefficients altering this relationships what is used to predict the quality with no reference of the original image.

Some metrics defined under this approach take the objective quality assessment as an information loss problem, using techniques related to information theory [123, 6]. In [6] authors propose to approach the quality assessment problem as an information fidelity problem, where a natural image source communicates with a receiver through a channel. The channel imposes limits on how much information can flow from the source (natural image), through the channel (distortion process) to the receiver (human observer). So they model the input and the output of the channel. The natural image is modeled using Gaussian Scale Mixtures (GSM) which have been reported as very appropriate to model the marginal density functions of the wavelet coefficients and the highly space-variant local statistics of a wavelet transformed natural image [127]. The distortion model is a simple attenuation and additive Gaussian noise model in each subband. Given the source and the distortion the Information Fidelity Criterion (IFC) is the mutual information between the source and the distorted image, i.e. the statistical information that is shared. An important feature of the IFC is that does not involve any parameters associated to display devices, data from psychophisical experiments, viewing configuration, or any stabilizing constants. The IFC is not a distortion metric, but a fidelity criterion, i.e. ranges from zero (no fidelity) to infitity (perfect fidelity).

1.5 Comparison of QAM

As previously mentioned, each QAM gets the quality of the image/video using their own and specific scale that depends on its design. Therefore this raw quality scores cannot be compared directly, even though the range of the values (the scale) is the same. In order to compare fairly the behavior of various metrics for a set of images or sequences, the objective quality index obtained from each metric has to be converted into a common scale.

When reviewing the performance comparisons that authors made in their QAM proposals, few details are provided about the comparison procedure itself. So it is difficult to replicate these results. In addition, different tests, with the same image set and even with the same subjects, can provide slightly varying results for a set of metrics, but as explained in [128] the results should be in line when test are correctly done.

In VQEG, subjective tests were repeated by several laboratories and the Pearson correlations between results by different laboratories range from 0.924 to 0.986 with mean of 0.97 confirming that even the best test methodologies cannot fully compensate for the uncertainty related to human factors such as test subjects and the consistency and interpretation of instructions. These results suggest also that slightly less consistent MOS scores are obtained in subjective tests carried out with image databases containing several different types of distortions than the obtained when the database has only a specific type of artifacts.

Authors in [128] reviewed the sources of inaccuracy of each step of the QAM comparing process shown at Fig. 1.30. Test video sequences or images from a set or database with known subjective scores (MOS or DMOS) are the input to the QAM. The QAM provides its quality indices or raw scores. Then, regression analysis is used to find a function that maps the obtained raw scores into subjective quality scores. Finally a correlation analysis is performed to estimate how accurately the subjective scores are predicted from the objective quality indices. The set of sequences or images in the database are called the metric “training set” because are used to fix the regression function.

The sources of inaccuracy in this process, may be related to many factors as the reliability of the subjective reference data, the types and degree of the distortions in the images or videos, the selection of the content that made up the training and testing sets and even the use and interpretation of the correlation indicators. This sources of inaccuracy can lead to quantitative differences when the same QAM is tested by different authors, even when the tests are correctly done.

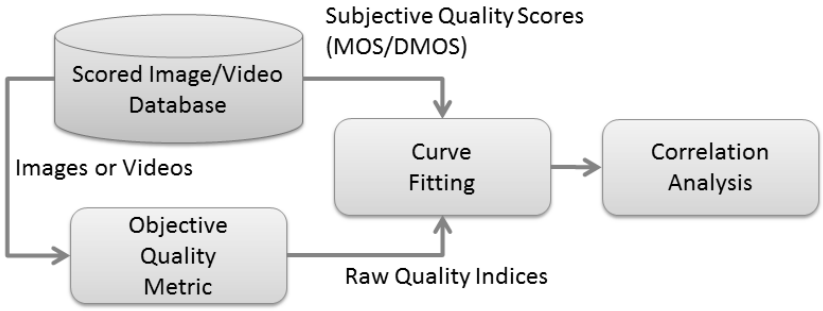


Figure 1.30: Block Diagram of the QAM evaluation process

The method in Fig. 1.30 is the one proposed by the VQEG [59] with some refinements proposed in other relevant comparison tests [129], where the target scale used is DMOS scale (Differences Mean Opinion Score). From a a subjective test, for example a Double Stimulus Continuous Quality Scale (DSCQS) method as suggested in [59], the Mean Opinion Score (MOS) can be calculated for the source and distorted versions of each image or sequence in this set. The scale used by the viewers goes from 0 to 100. These scores are converted into difference scores and processed further as explained in [6] to get the DMOS also in the 0-100 range.

The DMOS is the difference between the MOS value obtained for the original image/sequence and the MOS value obtained for the distorted one. So, for a particular image or sequence its DMOS value gives the mean subjective value of the difference between the original and the distorted versions. A value of 0 means no subjective difference found between the images by all the viewers. Due to the nature of the subjective test this value is very unlikely.

Performing a subjective test following the recommendations of the VQEG is not an easy and quick task, because a lot of technical requirements must be taken into account and some statistical analysis must be done to the raw subjective data in order to follow VQEG recommendations [58]. So as exposed in figure 1.30 the source of the subjective scores for such comparison test, is usually an image or video database with the associated MOS or DMOS values.

In [71], author review a set of perceptually scored image databases, LIVE [130], CSIQ [131], IVC [132], Toyama [133], A57 [134], TID [135] and WIQ [136]. In addition some video databases as CSIQ [137], TUM [138], LIVE[15], VQEG-FR-PhaseI [139] and VQEG-HDTV-PhaseI [140] also include subjective values. For the majority of the databases analyzed in [71] results are in accordance

with the results of our tests which are shown bellow.

1.5.1 Metric Comparison Results

The issues summarized in [128] encouraged and guided us to perform our own comparison test with a set of the most relevant QAM whose source code or test software has been made available by their authors. The results of our tests, as expected, were slightly different from other comparison tests but remain in line with their results as [128] predicts. The metrics used in our study are summarized herewith.

- The DMOSp-PSNR metric. We translate the traditional PSNR to the DMOS space applying a scale-conversion process. We call the resulting metric DMOSp-PSNR.
- The Mean Structural SIMilarity index [110] (MSSIM) from the Structural Distortion/Similarity Framework. In the reference paper, this FR metric was tested against JPEG and JPEG2000 distortion types. We test its performance with the new distortion types available in the second release of Live Database, "Live2 Database" since it is considered a generalist metric.
- The Visual Information Fidelity (VIF) metric [141] from the Statistics of Natural Images Framework. A FR metric that quantifies the information available in the reference image, and determine how much of this reference information can be extracted from the distorted image.
- The No-Reference JPEG2000 Quality Assessment (NRJPEG2000) [117] from the Statistics of Natural Images Framework. A NR metric that uses Natural Scene Statistical models in the wavelet domain and uses the Kullback-Leibler distance between the marginal probability distributions of wavelet coefficients of the reference and distorted images as a measure of image distortion.
- Reduced-Reference Image Quality Assessment (RRIQA) [123] from the Statistics of Natural Images Framework. The only RR metric under study. It is based on a Natural Image Statistical model in the wavelet transform domain.
- The No-Reference JPEG Quality Score (NRJPEGQS)[101] from the HVS Properties Framework. A NR metric designed specifically for JPEG compressed images
- The Video Quality Metric[94] (VQM General Model) from the HVS Properties Framework. The VQM uses RR parameters sent through an ancillary channel

that requires at least a 14% of the uncompressed sequence bandwidth. Although being conceptually a RR metric, it was submitted to the VQEG FR-TV test because the ancillary channel can be used to receive more detailed and complete references from the original frames, even the original frames themselves.

As exposed, the first step in the comparison method is to perform a subjective test to get the DMOS values. We have not done such a subjective test. Instead of this, we have use directly the DMOS values published in the Live Database Release 2 [130] and in the VQEG Phase I Database [139] following the method shown in Fig. 1.30. Image metrics were applied to each frame of the sequences and the mean raw value for all the frames was translated to the DMOSp scale.

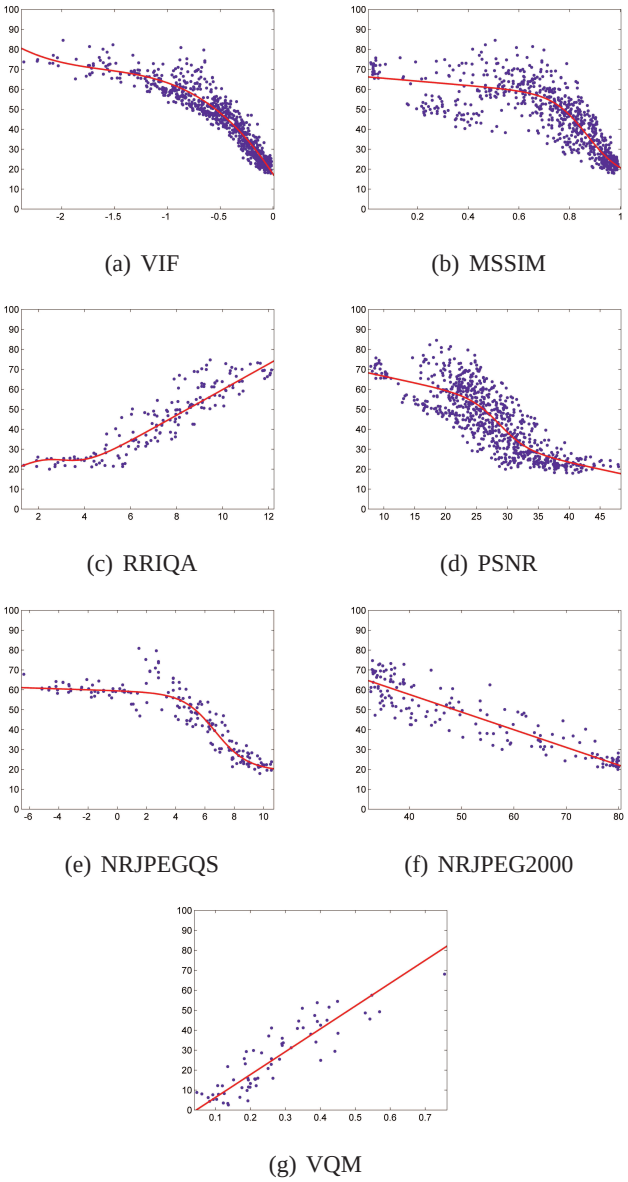
As suggested in [128, 142] the performance evaluation of the metrics (Correlation Analysis step) should be computed after a non-linear curve fitting process. A linear mapping function cannot be used because quality scores are rarely scaled uniformly in the DMOS scale, because different subjects may interpret vocabulary and intervals of the rating scale differently, depending on the language, viewing instructions and individual psychological characteristics. Therefore a linear mapping function would give too pessimistic view of the metric performance. Several mapping functions could be selected for this purpose, such as cubic, logistic, exponential and power functions, being monotonicity the main property that the function must comply with, at least in the relevant range of values.

The non-linear mapping function between the objective and the subjective scores used in our tests, was the one suggested by the VQEG and other relevant authors [58, 59, 129], and is shown in Equation 1.4. It is a parametric function that converts the metric raw score into a value in a Predicted DMOS (DMOSp) scale. In this DMOSp scale the quality score given by a metric for a specific image/sequence is directly comparable with the one given by the other metrics for the same image/sequence.

$$Quality(x) = \beta_1 logistic(\beta_2, (x - \beta_3)) + \beta_4 x + \beta_5 \tag{1.4}$$

$$logistic(\tau, x) = \frac{1}{2} - \frac{1}{1 + \exp(\tau x)} \tag{1.5}$$

Equation 1.4 has five parameters, from β_1 to β_5, that are fixed by the curve fitting process. We have not found in the literature any mapping function jointly with the parameter values for any image/video database. So, we have calculated



(a) VIF (b) MSSIM

(c) RRIQA (d) PSNR

(e) NRJPEGQS (f) NRJPEG2000

(g) VQM

Figure 1.31: Dispersion plots of the evaluated metrics including the curve fit for Eq. 1.4

these parameters based on sets of images and sequences that conforms our "training set".

In Fig. 1.31(a) to Fig. 1.31(g) the dispersion plots used in our fitting process, for all the selected metrics are shown. Each point of the scatter-plots corresponds

to an image in the training set and represents the DMOS value obtained from the scores given by a set of viewers.

The X-axis of the plots correspond to the raw values given by each of the metrics. In the Y-axis we have the corresponding DMOS values from the database. The curve fitting process gives us the parameters for Equation 1.4, which is represented by the solid curves. Depending on the metric, increasing x-axis values can have different interpretations, for example, in Fig. 1.31(a) for the VIF metric, 0 corresponds to the highest quality reported by the metric and decreasing values means lower quality, whereas in Fig. 1.31(b) for the MSSIM metric a value of 0 in the X-axis corresponds to lowest quality value being 1 the corresponding value to best reported quality.

The quality of the images in the subjective test is variable, covering a large range of distortion types and intensities for each distortion. Image distortions go from very highly distorted to practically undistorted ones. The viewers gave their scores for each image in the set, obtaining the average DMOS value. As shown in Fig. 1.31(a), the dynamic range of the average DMOS values does not reach the limits of the DMOS scale (0 and 100) for any distortion type, therefore the fitted curve predicts DMOSp values inside the same dynamic range. This is the reason why for a raw score of 0 (the best possible quality for the metric in this case) the predicted DMOSp value is not 0, i.e. there was no image scored with a DMOS value of 0, instead of that, the best DMOSp value obtained is around the value of 20. So, in the case of the VIF metric its dynamic DMOSp range varies from 20 to 80. The rest of the metrics have slightly different dynamic DMOSp ranges because the set of images used in each case is different, as we explain below.

Table 1.1: Equation parameters of metrics under study

	β_1	β_2	β_3	β_4	β_5
MSSIM	-39.5158	14.9435	0.8684	-10.8913	46.4555
VIF	-3607.3040	-0.5197	-1.6034	-476.0144	-693.3585
NRJPEGQS	37.6531	-0.9171	6.6930	-0.2354	40.7253
NRJPEG2000	37.3923	0.8190	0.6011	-0.8882	74.5031
RRIQA	-18.9995	1.5041	3.0368	6.4301	5.0446
PSNR-DMOSp	23.2897	-0.4282	28.7096	-0.6657	61.5160
VQM-GM	-163.6308	6.3746	-7.6192	114.4685	76.6525

Once the beta parameters have been obtained for each metric (see Table 1.1) the raw scores can be translate to the DMOSp scale shared by all metrics and hence, we can compare the results given by different metrics while scoring the same image.

The fidelity to subjective scores of a metric is consider high if the (PCC) and the Spearman Rank Order Correlation Coefficient (SROCC) are close to 1 and the Outlier Ratio (OR) is low [70]. In table 1.2 the performance parameters of our fittings are shown. These performance parameters show the degree of correlation between the DMOSp values and the subjective DMOS values provided by the viewers. Performance validation parameters are the PCC the Root Mean Squared Error (RMSE), the SROCC and the OR. In table 1.3 we include also the Mean, Max and Standard Deviation of error. In order to interpret correctly the meaning of "error" is worth to remember that the resulting DMOSp values for each metric correspond to values located in the fitted curve plotted in red in figures 1.31(a) to 1.31(g). So the error for each DMOS point (blue points) is the distance (absolute value) to the fitted curve. Outliers have not been removed from sets for obtaining these error parameters who provide an idea of how sparse or close to the fitted curve are the cloud of points in each case.

- The PCC is the linear correlation coefficient between the Prredicted DMOS (DMOSp) and the subjective DMOS. It measures the prediction accuracy of a metric, i.e., the ability to predict the subjective quality ratings with low error.
- The SROCC is the correlation coefficient between the DMOSp and the subjective DMOS. It 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 quality ratings.
- OR is defined as the percentage of the number of predictions outside the range of 1.5 times the standard deviation of the subjective results. It measures the prediction consistency, i.e., the degree to which the metric maintains the prediction accuracy.
- Mean Error is the mean of the errors produced when obtaining each DMOSp value in relation to their original DMOS value (for all images in the used "training set").
- Max Error is the highest error produced when obtaining the DMOSp values.
- Std Error is the Standard Deviation of errors

Another key point to consider while comparing QAM [128] is the correct selection of the image or video sequence sets used as "training set". The "training set" is used to perform the curve fitting process. This set should be chosen with special care and must be excluded from validation tests. So for each metric, the fitting process must be done using images or sequences with impairments that the

Table 1.2: Statistical parameters of the goodness of fit

	PCC	RMSE	SROCC	OR
MSSIM	0.8625	8.1809	0.8510	0.0359
VIF	0.9502	5.0187	0.9528	0.0282
NRJPEGQS	0.9360	5.7006	0.9020	0.0455
NRJPEG2000	0.9099	6.7306	0.9021	0.0059
RRIQA	0.9175	6.5393	0.9194	0.0353
PSNR-DMOSp	0.8257	9.0852	0.8197	0.0064
VQM-GM	0.8957	7.6435	0.9021	0.0000

Table 1.3: Error related parameters of the goodness of fit

	Mean Err	Max Err	Std Err
MSSIM	6.2130	24.3351	8.1792
VIF	3.8676	25.4201	5.0219
NRJPEGQS	3.9946	21.9940	5.6562
NRJPEG2000	5.4029	18.4913	6.7506
RRIQA	4.8190	19.2447	6.4961
PSNR-DMOSp	7.2712	24.7603	9.0911
VQM-GM	6.3009	16.4353	7.6897

metric is designed to handle. See [128] for details of how an incorrect selection of the image “training set” can influence in the final interpretation of the statistics used in the correlation analysis.

So, the MSSIM, VIF, RRIQA and DMOSp-PSNR metrics were “trained” with the whole Live2 Database because they are intended to be generalist metrics. The NRJPEGQS was “trained” only with the JPEG distorted images of Live2 database as this metric is designed only to handle this type of distortions. And for the same reason the NRJPEG2000 was “trained” only with the JP2K distorted images of the Live2 database and the VQM-GM was “trained” with a subset of 8 video sequences and its 9 corresponding HRCs of the VQEG Phase I database in a bitrate range of 1 to 4Mb/s.

It is important to mention that each of these “training sets” have different dynamic ranges in the DMOS scale as the degree of distortions applied to the images in each set is different.

We define as “homogeneous metrics” those which were trained with the same sets and therefore sharing the same DMOS dynamic range. So, metrics are called to be “heterogeneous metrics” when they were trained with different sets.

In our study all the metrics have been “trained” only with the luminance information and as suggested, only appropriate impairments are used while conforming the “testing sets” for each metric. .

From the performance results we can conclude than with the images and sequences that comprise our training sets the QAM that best performance gives, i.e. a higher correlation with subjective results, is the VIF metric.

1.5.2 Analyzing Metrics Behavior

In this section we are interested in analyze the metrics behavior when measuring image and video distortions produced in 1) compression scenarios at different rates and 2) distortions produced by packet losses in mobile ad-hoc network scenarios with variable degrees of network congestion an node mobility.

In Compression Environments

In this section we will study the behavior of the QAM under evaluation when assessing the quality of compressed images and sequences with different encoders. As exposed before, in the development of a new encoder or when performing modifications to existing ones, the performance of the proposals must be evaluated in terms of perceived quality by means of the R/D behavior of each encoder. The distortion metric commonly used in the R/D comparisons is PSNR.

So, in this test environment, we will work with the selected metrics as candidates to replace the PSNR as the quality metric in a R/D comparison of different video codecs. In this case, we will use a set of video encoders and video sequences in order to create distorted sequences Hypothetical Reference Circuit (HRC) at different bitrates, and analyze the results of the different QAM under study. Also, we will consider the metric complexity in order to determine their scope of application. For the tests we have used an Intel Pentium 4 CPU Dual Core 3.00 GHz with 1 Gbyte RAM. The programming environment used is Matlab 6.5 Rel.13. The fitting between objective metric values and subjective DMOS scores was done using the Matlab curve fitting toolbox looking for the best fit in each case. The codecs under test are:

- H.264/AVC [143]
- Motion-JPEG2000 [144]
- Motion-LTW [145]

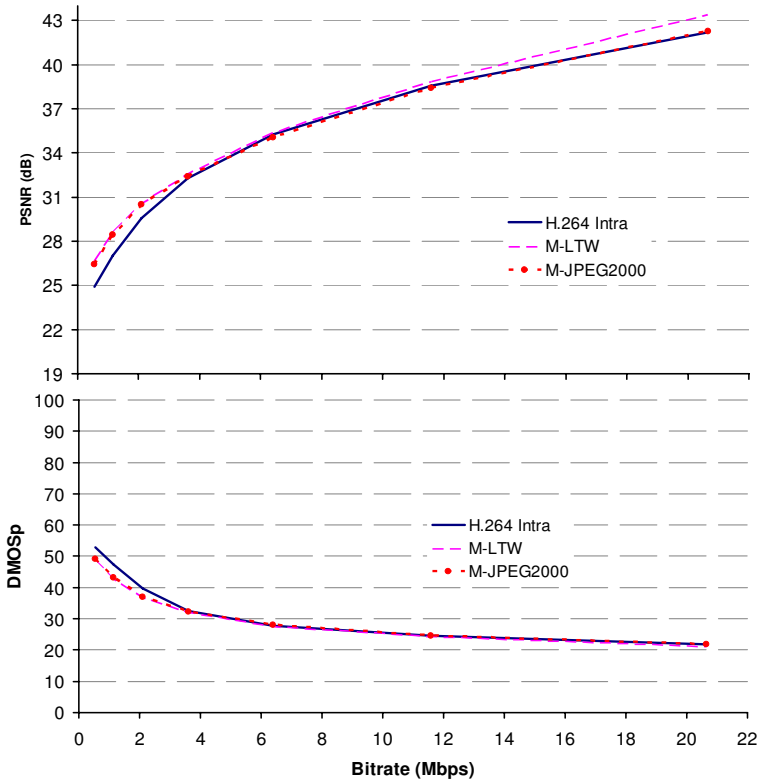


Figure 1.32: PSNR vs DMOSp-PSNR for the evaluated codecs (mobile sequence)

A R/D plot of the different video codecs under test, using the traditional PSNR as a distortion measure, is shown in the upper panel of Fig. 1.32. It is usual to evaluate performance of video codecs in a PSNR range varying from 25-27 dB to 38-40 dB, because it is difficult determine which one is better with PSNR values above 40 dB.

We convert the traditional PSNR to a metric that we call DMOSp-PSNR by applying the scale-conversion process explained in section 1.5. We can consider the DMOSp-PSNR metric to be the "*subjective*" counterpart of the traditional PSNR. It is the same metric, though expressed in a different scale. The DMOSp scale denotes distortion, thereby quality increases as DMOSp value decreases. The main difference between PSNR and its counterpart DMOSp-PSNR is that the saturation effect is fixed, as we can see in the lower panel at Fig. 1.32. As the only modification that has been done to the PSNR metric is the mapping process with the DMOS data, the raw values of the PSNR do not change, therefore DMOSp-PSNR metric does not fix the known drawbacks shown in Fig. 1.2.

This saturation effect, at high qualities, is not captured by the traditional PSNR that increases steadily as the bitrate rises, as shown in the upper panel of Fig. 1.32. Subjective saturation effect is noticeable above a specific quality value (saturation threshold) where the DMOSp values practically do not change. In our tests the saturation threshold were located at a bitrate of 11.58 Mbps. This behavior is repeated for all the evaluated codecs and video formats, confirming that there is no noticeable subjective difference when watching the sequences at the two highest evaluated bitrates (11.58 and 20.65 Mbps).

For each bitrate value below the saturation threshold the DMOSp-PSNR metric arranges the codecs (by quality) in the same order as the PSNR does, as expected, because in fact it is the same metric. This quality sorting, below the saturation threshold, agrees also with the results of the subjective tests that we performed (see below), and this behavior is repeated for all the evaluated sequences and bitrates.

Since PSNR, and therefore DMOSp-PSNR, are known to be inaccurate perceptual metrics for image or video quality assessment, we analyze the remaining metrics under study for all codecs and bitrates. From section 1.5 we know that the expected behavior of a QAM when scoring an image or sequence at different bitrates shoud be:

- For bitrate values below the saturation point, it should give a decreasing quality value as the bitrate decreases.
- For bitrate values above the saturation point, the perceptual quality value should be almost the same.

So, we run all the metrics for each HRC (sequence and codec) and analyzed the resulting data between consecutive bitrates, obtaining the quality scores in the DMOSp space. Then, a simple subjective DSCQS test was performed with 23 viewers in order to detect if there was or not perceptual differences at high bitrates, i.e. above the saturation threshold, for the tested sequences. For each sequence and encoder, the three HRCs with higher bitrates were presented to the viewers, each time in a different order, so that viewers did not know the rate for each sequence. These HRCs were: the first one located below saturation point (6.4 Mbps) and the two located in the saturation region. For example in figure 1.32 this three points are located at 6.4 Mbps (below threshold) and the two rightmost points at 11.58 and 20.65 Mbps. The test shows that:

- All the viewers detected some perceptual differences bellow threshold.

- No perceptual differences were detected above saturation threshold.

- Above saturation threshold, the DMOSp differences for the tested HRCs vary from 0.37 to 6.73 DMOSp points depending on the metric, sequence and encoder. See the whole set of values in tables tab. 1.6 to tab 1.11 at the end of the chapter.

So, from the results of our subjective test, we can initially conclude that above saturation differences up to 6.73 DMOSp values are perceptually indistinguishable.

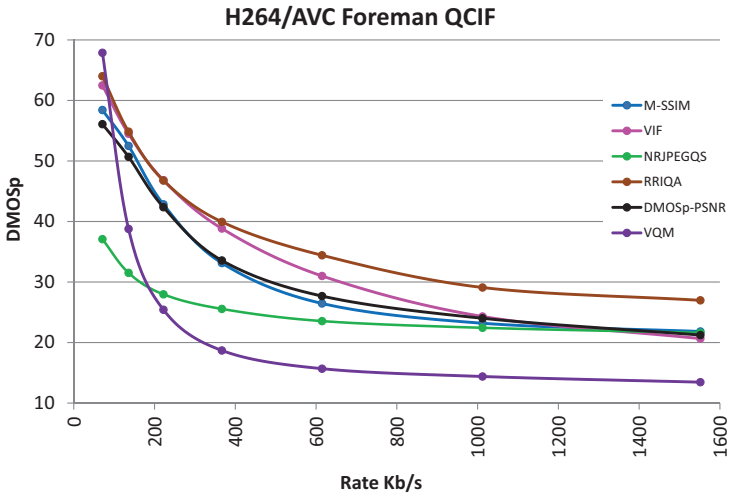
In figure 1.33 we can see examples of the R/D plots used for comparing the metrics. Each of these figures, show the resulting DMOSp R/D curves for all the metrics when applied to the same sequence and encoder at different bitrates. More figures are shown at the end of this chapter, in section 1.5.4. As shown, in both examples of figure 1.33, the perceptual saturation effect is captured by all the QAM at high bitrates (high quality) regardless of the encoder. The same holds for all the sequences and encoders.

Some metrics are missing in each of the example plots in figure 1.33. In the upper plot, the HRCs were encoded with the H.264/AVC codec, and therefore the NRJPEG2000 metric is omitted because it is not designed to handle DCT transform distortions. In the same way, in the bottom plot, where HRCs were encoded with M-JPEG2000, the NRJPEGQS metric is omitted because it is not designed to handle the distortions related to the Wavelet transform.

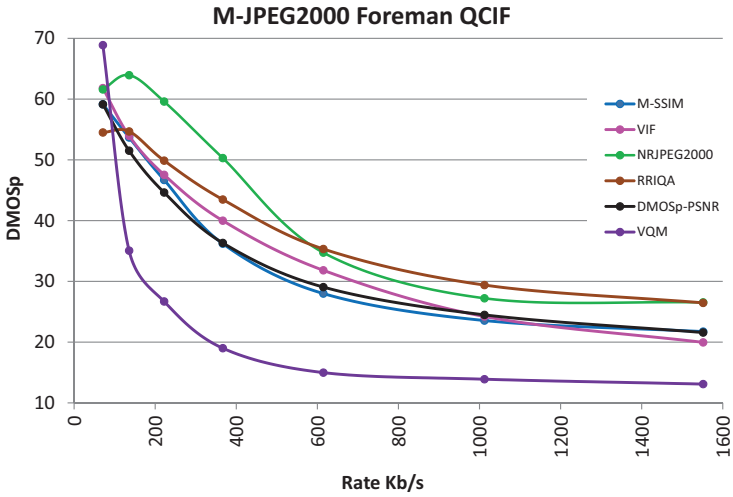
As mentioned in section 1.5, monotonicity is expected in the mapping function. So, the expected behavior of the metrics should also be monotonic, i.e. metrics should indicate lower quality values as the bitrates decreases. However, if we look at the lower plot of Fig. 1.33, and focusing this time on the two lowest bitrates, the quality score given by both, the RRIQA and NRJPEG2000 metrics, increases as the bitrate value decreases. This behavior is contrary to the expected one for a QAM. Remember that lower values of DMOSp represent better perceptual quality. More figures with the same behavior can be found in section 1.5.4 at the end of this chapter.

To illustrate this behavior, in Fig. 1.34 we show the first frame of the Foreman sequence at these bitrates (for the QCIF frame size). The left images is encoded at 70 Kbps, and the right image at 135 Kbps. After a visual comparison, the right image receives a better subjective score than the left one though the mentioned metrics state just the opposite in this particular case.

Our results for the compression environment stated that:



(a)



(b)

Figure 1.33: QAM comparison using the same sequence with different codecs (a) H264/AVC Intra (b) M-JPEG2000

- NRJPEG2000 offers wrong quality scores between the two highest compression ratios with the M-JPEG2000 codec, for QCIF and CIF sequences.
- RRIQA also failed with this NRJPEG2000 at high compression ratios, but only for small the Foreman QCIF sequence.
- All the other metrics exhibit a monotonic behavior for all bitrates regardless of the encoder and sequence being tested.

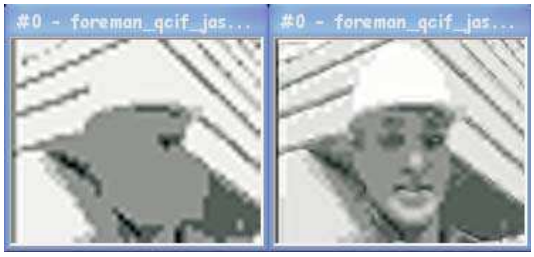


Figure 1.34: First frame of Foreman QCIF encoded at 70 Kbps (left) and 135 Kbps (right)

Figure 1.33 will also help us to illustrate what was exposed in section 1.5, heterogeneous metrics should not be compared directly, because the dynamic range of the subjective quality scores in each training set is different.

Looking at upper plot in fig. 1.33 and focusing this time on the lowest bitrate, the DMOSp rating differences between metrics arrive surprisingly up to 30.79 DMOSp units. As the test sequence at this rate is the same for all metrics, this difference seems to be too high and lead us to think that something must be wrong here. In addition, there are three different behaviors or trends in the R/D curves. So, let us analyze what is wrong here.

The three different trends in fig. 1.33 correspond to the use of three different training sets. A exposed previously, VQM-GM was trained with VQEG sequences, NRJPEGQS was trained only with the JPEG distorted images, and the rest of the metrics trained with the whole set of distorted images in the Live2 database. Each trend is the result of a curve fitting process with different betas (parameteres) and these betas are directly dependent of the used training set (the set of distorted images presented to the viewers). This is the reason why the trends and slopes of the metrics below the saturation threshold are different and as shown are “grouped” together in both examples shown in figure 1.33.

So, when including in the same R/D plot, curves from different metrics, it would be preferable that they are homogeneous, and if not, this fact must be told in order to avoid misleading conclusions about the compared performance

between heterogeneous metrics. R/D plots with heterogeneous metrics should not be use to determine which metric is the best, not even R/D plots with only homogeneous metrics. These type of plots are useful however, to analyze the behavior of the metrics for each encoder and/or sequence, to compare and measure differences in quality among metrics while coding at the same rates and to detect some anomalous behaviors, as the ones presented above.

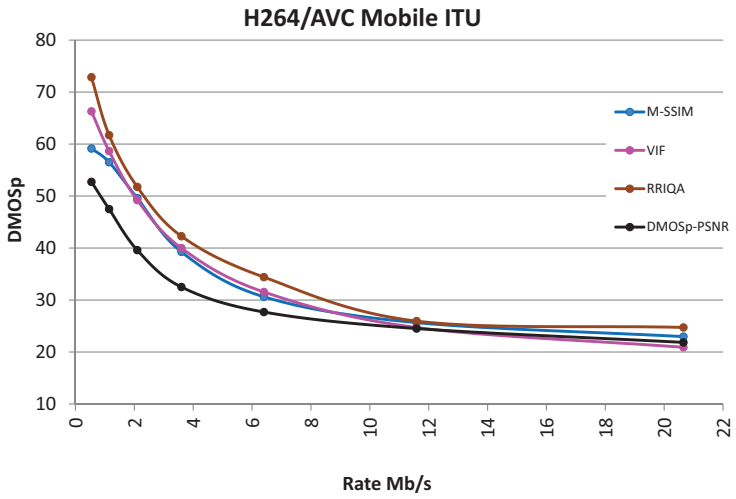


Figure 1.35: QAM comparison plot with homogeneous metrics

In figure 1.35 only homogeneous metrics are shown. The trend of all the R/D curves is the same. Only by inspecting the curves, and comparing the QAM behavior in the bitrate range, it can not concluded which metric is the best. Is it the one with better DMOSp for all the bit-rate range? What if this metric is wrongly overrating the quality given by the observers?

Determining how good a metric works at a specific rate or for a bit-rate range, depends on how good the metric predicts the subjective scores given by human viewers, i.e. the best metric is the one who best mimics the human rates. This information is obtained from parameters like those of tables 1.2 and 1.3.

Our metric performance validation tests data tells that the VIF metric is the one which best fits the subjective DMOS values among the metrics in the same “training set”. So in plots, such as those from figure 1.33, the best performing metric can act as reference. Then, we can compare for each sequence and encoder how far from the reference the rest of the metrics are. Remember that not all the metrics can be use to score all the encoders, the metric should be able to handle the encoder specific produced distortions.

Table 1.4: Sequences included in the “test set”

Sequence	Frame	F.Num.	F.Rate
Foreman	QCIF: 176 x 144	300	30 fps.
Container			
Foreman	CIF: 352 x 288		
Container			
Mobile	640 x 512	40	

Once we have compared and analyzed the behaviors of the metrics and having chosen the best correlated to human perception one, we proceed with the encoder comparison. For this comparison, our “test set” comprise different standard video sequences commonly used in video coding evaluation as shown in table 1.4, using only the luminance component. We perform this test for each QAM being evaluated.

Fig. 1.36 represents an example of one of the R/D plots used for comparing the performance of the encoders being tested. In this case the plot shows how the VIF metric evaluate the performance of the encoders. In figures from fig 1.61 to fig 1.95 the rest of the metrics plots are shown.

For metrics “trained” with the same set, the ranking order of the encoders at a specific bitrate, should agree among metrics, and also with the subjective ranking given by the viewers. To check this, we performed a simple subjective test with 23 viewers in order to evaluate if we can trust the codec ranking order given by each metric, i.e. at a specific bitrate the metric order the encoders by quality in the same perceptual order that subjective one.

For each rate and sequence the reconstructed sequence of each encoder were presented simultaneously to the subjects. The ordering of the three sequences varies for each HRC, so that the subjects did not know which encoder correspond to each sequence. The subjects ranked the sequences by perceptual quality, if no differences were detected between pairs of sequences, they annotated this fact. After analyzing the viewers scores and removing outliers, the test confirms that the ranking order was consistent among homogeneous metrics, agreeing also with the subjective ranking.

In cases where viewers scored no subjective difference between two sequences, the metrics still gave slightly different values between encoders, being these differences in a range lower than 2.9 DMOSp units. When these differences

between metric values were higher, for example 3.11 DMOSp units at 2.1 Mb/s between H264/AVC and M-JPEG2000 in figure 1.36, most of the viewers could see some perceptual differences between the sequences, since they ranked H264/AVC to have better perceptual quality than M-JPEG2000 and M-LTW.

In order to determine how much difference, expressed in the DMOSp scale, is perceptually detectable, deeper subjective tests and research must be done, because from our studies, we already detect that the perceptual meaning of these DMOSp differences depend on the point in the DMOSp scale where we are working on. For example, for high quality (as stated before), DMOSp value differences up to 6.73 DMOSp points were imperceptible, however, at lower quality levels smaller differences (3.11 DMOSp points) were perceived.

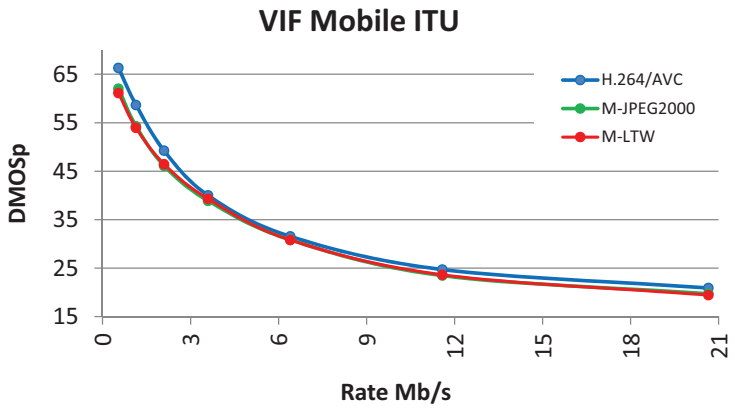


Figure 1.36: R/D performance evaluation of the three video codecs using Mobile ITU video sequence by means of VIF metric

Table 1.5: QAM Average scoring times (seconds) at frame and sequence level.

	QCIF		CIF		640 x 512	
	Frame	Seq	Frame	Seq	Frame	Seq
MSSIM	0.028	8.4	0.147	44.1	0.764	30.5
VIF	0.347	104.1	1.522	456.5	6.198	247.9
NRJPEGQS	0.01	3	0.049	14.6	0.201	8.1
NRJPEG2000	0.163	48.9	0.486	145.9	1.595	63.8
RRIQA(f.e.)	4.779	1433.7	6.95	2084.9	10.111	404.5
RRIQA(eval.)	0.201	60.2	0.635	190.6	2.535	101.4
DMOSp-PSNR	0.001	0.3	0.006	1.7	0.02	0.8
VQM-GM	0.023	6.975	0.093	27.900	0.300	12.024

Finally, Table 1.5 shows, grouped by frame sizes, the mean frame evaluation time and the evaluation time for the whole sequence that each metric spent to assess its raw quality value.

In the test, we have disaggregated the time spent in performing the quality comparison from other times spent in performing other steps, for some metrics. This way we can compare times jointly or in a separate manner. For example, times spent in the two steps of RRIQA, features extraction (f.e.) and quality evaluation (eval.), have been separately measured.

So for example if we do not take into account calibration and color conversion times when comparing against the VQM-GM, for CIF sequences the VQM-GM is faster than the other metrics, except NRJPEGQS and DMOSp-PSNR.

DMOSp-PSNR is the less computationally expensive metric for all frame sizes. On the other hand, RRIQA and VIF are the slowest metrics (as they run the sterable-pyramid, a linear multi-scale, multi-orientation image decomposition).

In MANET environments

Our objective in this section is to analyze the behavior of the candidate metrics in the presence of packet losses under different MANET scenarios. In order to model the packet losses in these error prone scenarios, we use a three-state Hidden Markov Model (HMM) and the methodology presented in [146]. HMMs are well known for their effectiveness in modeling bursty behavior, relatively easy configuration, quick execution times and general applicability. So, we consider that they fit our purpose of accelerating the evaluation process of QAM for video delivery applications on MANET scenarios, while offering similar results to the ones obtained by means of simulation or real-life testbeds. Basically, by the use of the HMM, we define a packet loss model for MANET that accurately reproduces the packet losses occurring during a video delivery session.

The modeled MANET scenario is composed of 50 nodes moving in an 870x870 square meters area. Node mobility is based on the random way-point model, and speed is fixed at a constant value between 1 to 4 m/s. The routing protocol used is DSR. Every node is equipped with an IEEE 802.11g/e enabled interface, transmitting at the maximum rate of 54 Mbit/s up to a range of 250 meters. Notice that a QoS differentiated service is provided by IEEE 802.11e [147]. Concerning traffic, we have six sources of background traffic transmitting FTP/TCP traffic in the Best Effort MAC Access Category. The foreground traffic is composed by real traces of an H.264 video encoded (using the Foreman CIF

video test sequence) at a target rate of 1 Mbit/s. The video source is mapped to the Video MAC Access Category.

We apply the HMM described above to extract packet arrival/loss patterns for the simulation traces, and later replicate these patterns for testing. We describe two environments: (a) congestion related environment, and (b) mobility related environment.

The congestion environment is composed of 6 scenarios with increasing level of congestion, from 1 to 6 video sources. The mobility environment is composed of 3 scenarios with only one video source, but with increasing degrees of node mobility (from 1 to 4 m/s).

For each of these scenarios we get different packet loss patterns provided by the HMM that represents each scenario.

After an analysis of the packet losses, different patterns are defined:

- Isolated small bursts represent less than 7 consecutive lost packets. As each frame is split in 7 packets at source, isolated bursts will affect to 1 or 2 frames, but none of them will be completely lost. This error pattern is mainly due to network congestion scenarios, where some packets are discarded due to transitory high occupancy in the wireless channel or buffers at relaying nodes.
- Large packet loss bursts. Large Bursts cause the loss of one or more consecutive frames. Large packe error bursts are typically a consequence of high mobility scenarios, where the route to the destination node is lost and a new route discovery process shuld be started. This will keep the network link in down state during several seconds, losing a large number of consecutive packets.

We have used the H.264/AVC codec adjusting the error resilience parameters to the values proposed in [148], so that the decoder is able to reconstruct sequences even when large packet loss bursts occurs. H.264/AVC is configured to produce one I frame every 29 P frames, with no B frames and to split each frame in 7 slices, so we to put each slice into a separate packet and encapsulate its output in RTP packets. As suggested in [148], we also force 1/3 of the macroblocks of each frame to be randomly encoded in intra mode.

We have used the Foreman CIF seq. (300 frames at 30 fps) to build an extended video sequence by repeating the original one up to the desired video length. After running the encoder for each extended video sequence, we get RTP packet streams. We will apply them a packet erasure process, removing those packets declared lost by the HMM model. This process simulates packet losses

in the MANET scenarios, so a distorted bitstream will be delivered to the decoder. The decoder behavior depends on the packet loss burst type as follows:

- When an isolated small bursts appear, the decoder is able to apply error concealment mechanisms to repair the affected frames. The video quality decreases, and just after the burst, the reconstructed video quality recovers the quality by means of the random intra-coded macroblock updating. When the next I frame arrives, it completely stops error propagation.
- When the decoder faces large bursts, it stops decoding and waits until new packets arrive. This produces a sequence in the decoder that is shorter than the original one. Therefore, both sequences are not directly comparable with the QAM and so we freeze the last completely decoded frame until the burst ends.

Once we have comparable video sequences (original and decoded video sequences with the same length), we are able to run the QAM. Each metric produces an objective quality value for each frame in its own scale. Then, we perform the scale conversion to the DMOSp scale (see section 1.5).

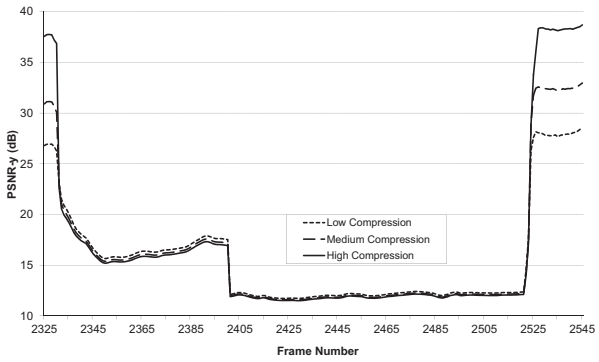


Figure 1.37: PSNR frame values during a long packet loss burst (from frame 2327 to 2525) at different bitrates.

Fig. 1.37 shows the objective quality value in the traditional PSNR scale at three different compression levels (Low compression, Medium compression and High compression) during a large packet loss burst. We observe the evolution of quality during the burst period. What the observer sees during this large burst is a frozen frame, with more or less quality depending on the compression level. The PSNR metric reports that quality drops drastically with the first frame affected by the burst, and decreasing even more as the difference between the frozen frame and the current frame increases. Nearly at the middle of the burst, an additional drop of quality can be observed. It corresponds to a scene change (with the

beginning of a new cycle of the foreman video sequence). At this point, the drastic scene change makes the differences between sequences even higher, and the PSNR metric scores with even worse values, reaching values as low as 10-12 dBs.

On the other hand, the perceived quality changes at these levels is quite difficult to evaluate. So, a better perceptually designed QAM should not score such a quality drop in this situation because quality saturates. When the burst ends, quality rapidly increases because of the arrival of packets belonging to the same frame number than the current one in the original sequence (frame 2525 in Fig. 1.37).

If during such a burst a QAM takes into account only the quality of the frozen frame, disregarding the differences with the original one (which changes over time), the effect of the burst would remain unnoticed for that metric, i.e. quality remain constant.

Fig. 1.38 shows the evolution of the candidate QAM during a large burst (similar to Fig. 1.37 but in this case in the DMOSp space). There is a panel for each compression level: the upper panel corresponds to high compression, the central panel to middle compression and the bottom panel to low compression. We observe some interesting behaviors that we proceed to analyze.

From a perceptual point of view, quality must drop to a minimum when one or more frames are lost completely and should remain that way until the data flow is recovered. It should not matter if a scene change takes place inside the large burst. VIF and MSSIM behaves this way. At the point of the burst, where the scene change takes place, both the VIF and MSSIM metrics have almost reached their 'bad quality' threshold regardless of the compression level and therefore there is no substantial change in the reported quality. The drop of quality to the minimum at the beginning of the burst evidence the lost of whole frames.

NR metrics do not detect the presence of a frozen frame (by dropping the quality score) as expected because the quality given by these metrics remain at the level scored for the frozen frame during the burst duration. So, NR metrics could not detect the beginning of a large burst, since lost frames will be replaced with the last correctly decoded frame (frozen frame) and the reference frames are not available for comparison. However, NR metrics detect the end of such bursts. Fig. 1.39 will help us to explain this behavior, showing how reconstruction is done after a large burst. This figure shows the impairments produced when the large burst ends. Fig. 1.39(a) is the current frame, the one being transmitted. Fig. 1.39(b) is the frozen frame that was repeated during the burst duration. When the burst ends, the decoder progressively reconstruct the sequence using the intra

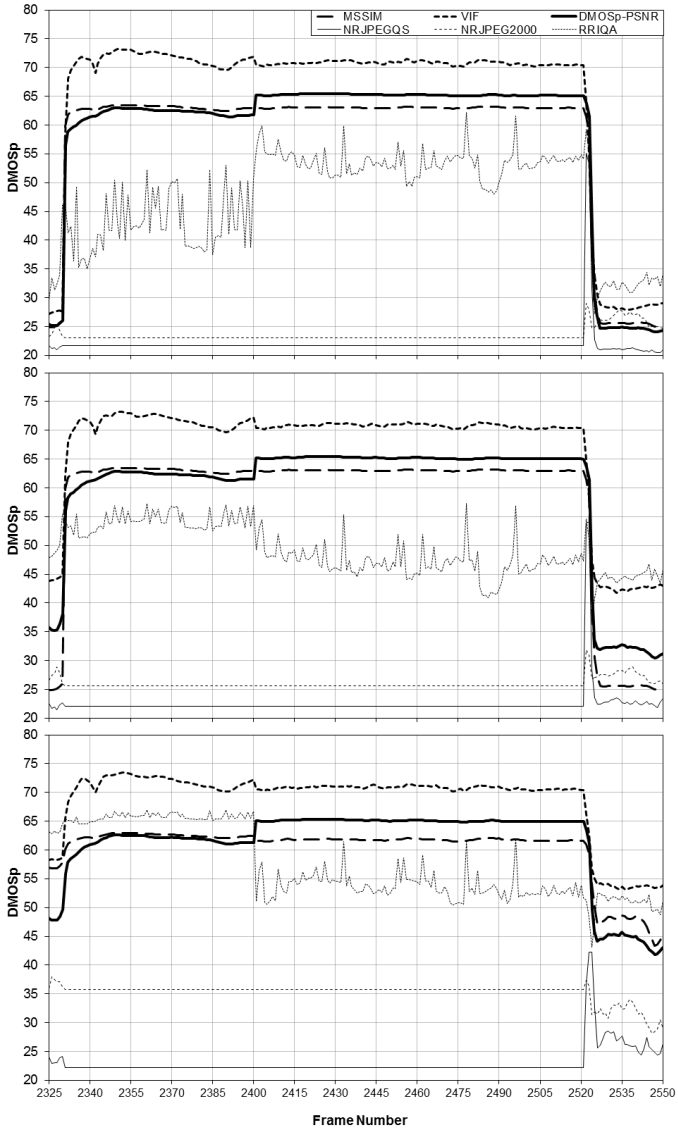


Figure 1.38: Metric comparison in the DMOSp space during a very large burst

macroblocks from the incoming video packets. So the decoder partially updates the frozen frame with the incoming intra macroblocks. This is shown in figures 1.39(c) and 1.39(d) where the face of the foreman appears gradually.

The gradual reconstruction of the frame with the incoming macroblocks is interpreted in a different way by NR metrics and FR metrics. When the

(a) (b) (c) (d)

Figure 1.39: Frame reconstruction after a large burst: (a)original frame, (b)last frozen frame, (c)(d)first and second reconstructed frames after the burst.

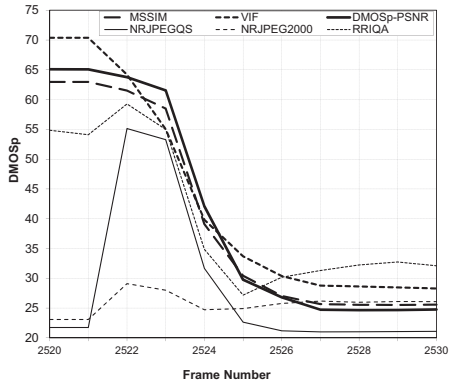


Figure 1.40: End of the large burst for the low compression panel. FR and NR metrics show the opposite behavior.

macroblocks begin to arrive, what happens at frame 2522 (see figure **??**) the NR metrics react scoring down quality, while the FR metrics begin to increase their quality score, just the opposite behavior. For a NR metric, without a reference frame, figure 1.39(c) has clearly worse quality than Fig. 1.39(b). But for a FR metric the corresponding macroblocks between Fig. 1.39(c) and Fig. 1.39(a) help to increase the scored quality.

So, NR metrics react only when the burst of lost packets affects frames partially, i.e. isolated bursts, and at the end of a large burst. The NRJPEGQS metric reacts harder (i.e it shows higher quality differences) than the NRJPEG2000 because it was designed to detect the blockiness introduced by the discrete cosine transform. When the frame is fully reconstructed then the score obtained with NR and FR metrics approaches again to the values achieved before the burst, which depends on the compression rate.

The RRIQA metric shows high variability in its scores between consecutive frames inside bursts. These variations become more evident as the degree of

compression decreases. The nature of the data sent through the ancillary channel, 18 scalar parameters obtained form the histogram of the wavelet subbands of the reference image, is very sensitive to loss of synchronism between the reference frame and the frozen one. On the decoder the same extracted parameters are statistically compared with the received through the ancillary channel. When this comparison is performed with two sets of parameters obtained from different frames, unexpected results appear.

Concerning the FR metrics, MSSIM, VIF and PSNR-DMOSp show a similar behavior or trend. MSSIM and PSNR-DMOSp show closer quality scores between them that the ones obtained with the VIF metric, which gives lower quality values than the other two metrics. This behavior is the same regardless the compression level inside the large burst. Leaving aside the PSNR-DMOSp, which is not really a QAM, the other two FR metrics (VIF and MSSIM) have the same behavior when facing large bursts.

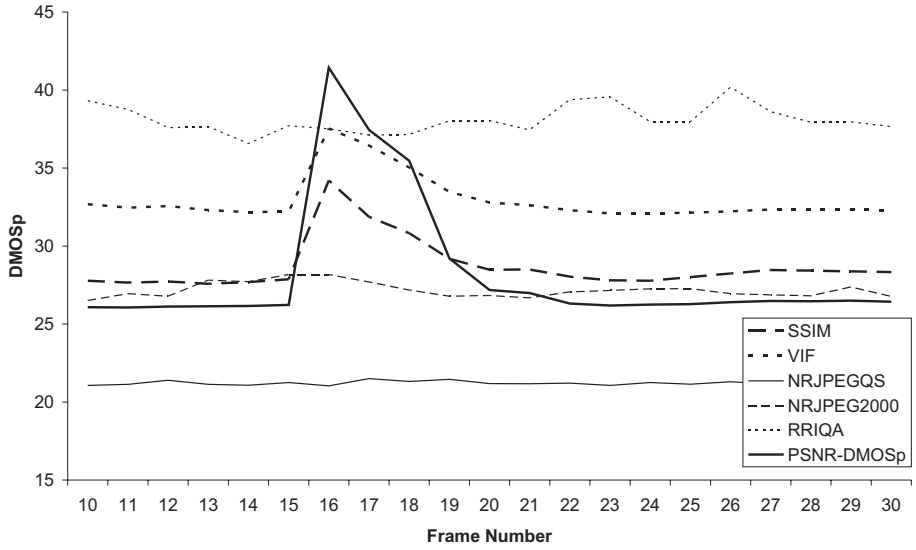


Figure 1.41: Metric comparison for an isolated burst

Fig. 1.41 shows an isolated burst. In this case, blur and edge shifting impairments are introduced altering only one frame. This fact is perceived only by the FR metrics and the NRJPEG2000, which is designed to detect this type of impairments. The error concealment mechanism of H.264/AVC needs up to 6 frames to achieve the same quality scores obtained before the burst. Fig. 1.42 shows the original frame (a) and three subsequent frames (b,c,d), where the effect of the lost packets is concealed by the H.264/AVC decoder.

(a) (b) (c) (d)

Figure 1.42: Packet loss affecting only one frame. (a) Original frame, (b,c,d) next three decoded frames

As defined previously, an isolated burst can affect one or two consecutive frames. In the last case, the behavior of the QAM when facing the isolated burst resembles the behavior of the metrics with a large burst. The difference is that the concealment mechanisms and the correct reception of part of the frames avoid a largest drop in the quality.

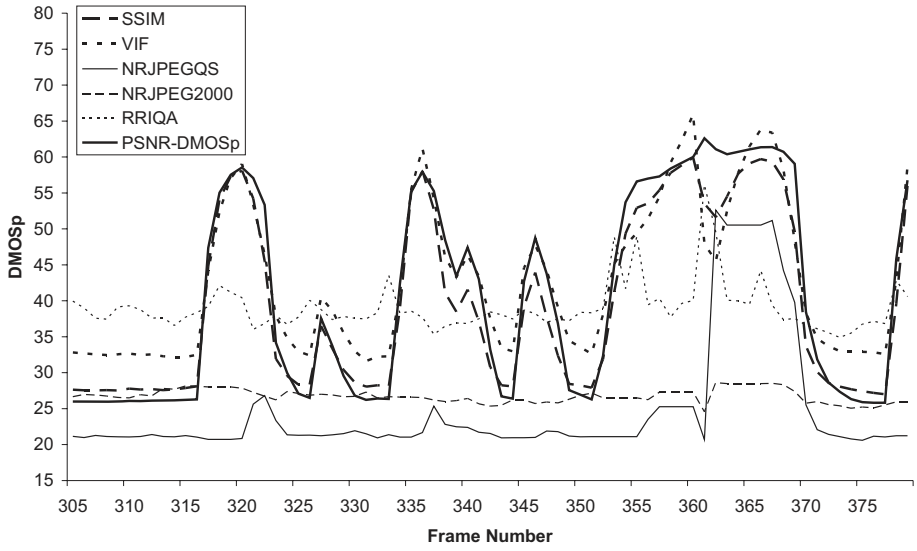


Figure 1.43: Frame interval where different type of bursts occurs consecutively.

Figure 1.43 shows multiple consecutive bursts (large and isolated) that behave as exposed previously. From left to right, we see a large burst followed by an isolated one. This pattern repeats again one more time, and at the right most part of the figure, between frames 352 and 372, two large bursts occurs consecutively, having a gap between them where new incoming packets arrive for a short period of time (frames 361 and 362).

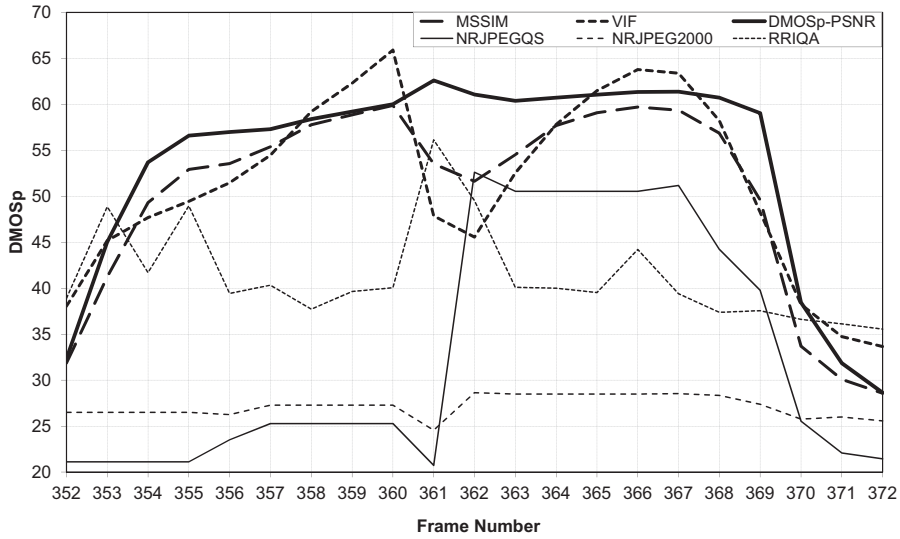
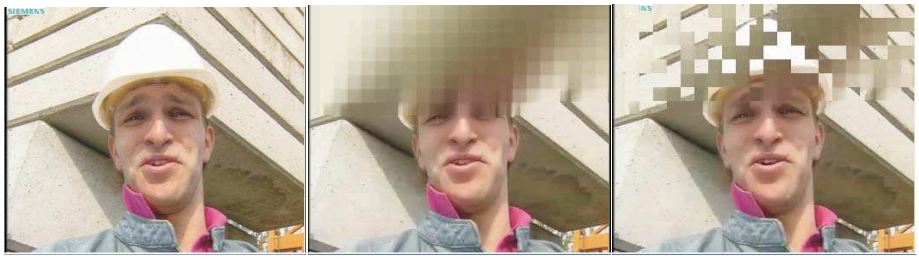


Figure 1.44: Detail from two consecutive long burst with incoming packets between them.



(a) (b) (c)

Figure 1.45: Decoded frames between two consecutive bursts, (a) original frame; Reconstructed frames (b) 361 and (c) 362

In Figure 1.44 we zoom into this area (frames 352 to 372) to analyze why the behavior of the DMOSp-PSNR metric differs from the other FR metrics during the gap between bursts. In the gap, the encoder is not able to reconstruct a whole frame because the gap is too small, i.e. between the two large burst only a small amount of packets arrive, and this is not enough to reconstruct a whole frame. So the involved frames (361 and 362) are partially reconstructed (figures 1.45(b) and 1.45 (c)). Both frames exhibit perfect correspondence in the lower half with the original one (Fig. 1.45(a)). Therefore, the scored quality must increase, at least to some extent, compared to the quality of the previous frozen frame, as occurs at the end of a large burst. This fact is only reflected by the VIF and MSSIM

metrics. The PSNR-DMOSp metric is not able to detect this because it is computed using information from the whole frame. For the VIF and the MSSIM, which are perceptually driven, the lower half of the frame increases their raw scores, in the same way as the human scores do. After frame 362, quality decreases again since the following frame is frozen too. So, VIF and MSSIM detect two consecutive loss burst while PSNR-DMOSp and the other metrics considers only a single larger one.

1.5.3 Conclusions

The main goal of this work was focused on looking for a Quality Assessment Metric that could be used instead of the PSNR when evaluating compressed video sequences with different encoder proposals at different bitrates, and to analyze the behavior of such metrics when compressed video is transmitted over error prone networks such as MANETs.

We explained the procedures that we followed to compare QAM metrics and alerted about some issues that arise when a comparison between heterogeneous metrics is made. The metrics must be compared using a common scale since the raw scores of the metrics are not directly comparable. The scale conversion process involves subjective tests and the use of mapping functions between the subjective MOS values and the metrics raw values. The parameters for the mapping function we used are provided. The metrics were first trained with a set of images from two open source image and video databases with known MOS values. The metrics were tested with another set of images and videos also taken from available databases. In order to perform a fair comparison, the training and testing sets used with each metric must use only impairments which the metric was designed to handle. We defined as heterogeneous metrics those that were trained with different sets of images or sequences. The R/D comparisons of heterogeneous metrics must be done carefully, focusing not only on the absolute quality scores, but also on the relative scoring between consecutive bitrates as the differences between DMOSp values are perceptually detected (or not) depending on the quality range. When metrics are trained with the same training set, differences in DMOSp values have the same perceptual meaning for all the metrics, but this may not be true between heterogeneous metrics. Normalizing the DMOSp scale when comparing heterogeneous metrics helps to detect these differences.

We performed the comparison between metrics in two environments: a compression environment and a packet loss environment. We performed several subjective tests in order to confirm that the analysis and the behavior of the

metrics was consistent with human perception. Our tests included the comparisons of three encoders by replacing the PSNR as distortion metric in their R/D curves with each of the candidate metrics.

From our results of the compression environment, we conclude that we can trust on the quality provided by the VIF metric metric, which is the one that obtains a better fit in terms of DMOS during the calibration process, and on how it ranks the performance of the tested encoders for the bitrate range under consideration. The NRJPEG2000 and the RRIQA metrics break monotonicity for very high compression levels when M-JPEG2000 is the evaluated encoder. For the rest of the bitrates, all the other metrics show a monotonic behavior for all the bitrate range and for all encoders.

The choice of a QAM to replace the traditional PSNR, when working in a compression framework with no packet losses, depends on the availability of the reference sequence. In applications where the reference sequence is not available, RRIQA is our choice because behaves similarly to FR metrics. If the reference sequence is available, the choice depends on the weight given to the trade-off between computational cost and accuracy. If time is the most important parameter, we will choose DMOSp-PSNR followed by VQM and MSSIM. If accuracy is more important, then the choice will be VIF and MSSIM metrics.

In the loss-prone environment, we analyzed the metrics behavior when measuring reconstructed video sequences encoded and delivered through error prone wireless networks, like MANETs. In order to obtain an accurate representation of delivery errors in MANETs, we adopted an HMM model able to represent different MANET scenarios.

The results of our analysis are the following: (a) NR metrics are not able to properly detect and measure the sharp quality drop due to the loss of several consecutive frames. (b) The RR metric has a non-deterministic behavior in the presence of packet losses, having difficulties ate identifying and measuring this effect when the video is encoded with moderate to high compression rates. (c) Concerning the other metrics, MSSIM, DMOSp-PSNR and VIF show a similar behavior in all cases. In summary, we consider that, although they exhibit slight differences in the Packet Loss framework, we propose the use of the MSSIM metric as a trade-off between a high quality measurement process (resembling human visual perception) and computational cost.

1.5.4 Figures and Tables

Table 1.6: Variation in DMOSp values between QAM above saturation point for the Foreman QCIF sequence

	H.264/AVC	M-JPEG2000	M-LTW	Max	Min
M-SSIM	1,36	1,82	1,79	1,82	1,36
VIF	3,65	4,26	4,13	4,26	3,65
NRJPEGQS	0,82			0,82	0,82
NRJPEG2000		0,68	1,21	1,21	0,68
RRIQA	2,12	2,93	2,31	2,93	2,12
DMOSp-PSNR	2,77	2,91	3,34	3,34	2,77
VQM	0,94	0,80	0,82	0,94	0,80
				4,26	0,68

Table 1.7: Variation in DMOSp values between QAM above saturation point for the Foreman CIF sequence

	H.264/AVC	M-JPEG2000	M-LTW	Max	Min
M-SSIM	1,84	2,38	3,32	3,32	1,84
VIF	4,18	3,96	4,91	4,91	3,96
NRJPEGQS	0,87			0,87	0,87
NRJPEG2000		0,82	2,43	2,43	0,82
RRIQA	2,72	2,93	2,03	2,93	2,03
DMOSp-PSNR	2,59	2,52	3,68	3,68	2,52
VQM	0,60	0,37	0,40	0,60	0,37
				4,91	0,37

Table 1.8: Variation in DMOSp values between QAM above saturation point for the Container QCIF sequence

	H.264/AVC	M-JPEG2000	M-LTW	Max	Min
M-SSIM	2,56	2,30	2,30	2,56	2,30
VIF	4,15	4,61	5,06	5,06	4,15
NRJPEGQS	0,90			0,90	0,90
NRJPEG2000		0,45	0,39	0,45	0,39
RRIQA	5,88	4,38	4,04	5,88	4,04
DMOSp-PSNR	2,61	2,66	3,02	3,02	2,61
VQM	1,96	1,88	0,45	1,96	0,45
				5,88	0,39

Table 1.9: Variation in DMOSp values between QAM above saturation point for the Container CIF sequence

	H.264/AVC	M-JPEG2000	M-LTW	Max	Min
M-SSIM	2,47	2,50	2,66	2,66	2,47
VIF	5,07	5,41	5,73	5,73	5,07
NRJPEGQS	0,88			0,88	0,88
NRJPEG2000		0,44	0,48	0,48	0,44
RRIQA	6,73	2,53	1,63	6,73	1,63
DMOSp-PSNR	2,67	2,49	2,90	2,90	2,49
VQM	1,06	0,69	1,14	1,14	0,69
				6,73	0,44

Table 1.10: Variation in DMOSp values between QAM above saturation point for the Moblie ITU sequence

	H.264/AVC	M-JPEG2000	M-LTW	Max	Min
M-SSIM	2,69	3,13	3,10	3,13	2,69
VIF	3,80	3,74	4,18	4,18	3,74
NRJPEGQS	1,45			1,45	1,45
NRJPEG2000		3,62	1,76	3,62	1,76
RRIQA	1,21	2,60	3,85	3,85	1,21
DMOSp-PSNR	2,66	2,84	3,28	3,28	2,66
VQM	0,71	0,81	1,20	1,20	0,71
				4,18	0,71

Table 1.11: Maximun and minimun variation in DMOSp values between QAM above saturation point for all the sequences

	Max	Min
Foreman qcif	4,26	0,68
Foreman cif	4,91	0,37
Container qcif	5,88	0,39
Container cif	6,73	0,44
Mobile itu	4,18	0,71
	6,73	0,37

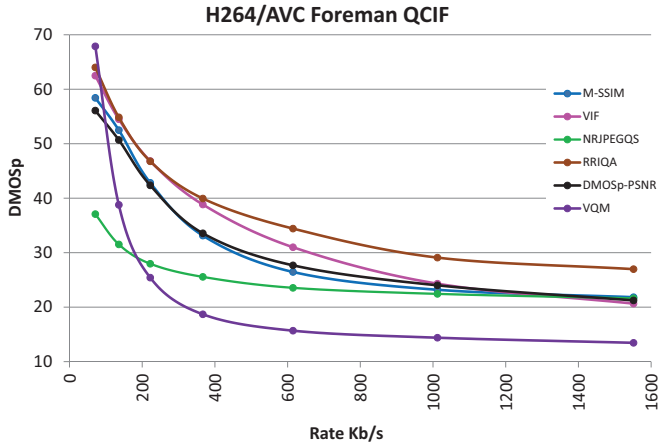


Figure 1.46: QAM comparison figures for Foreman QCIF and H264/AVC codec in Intra mode

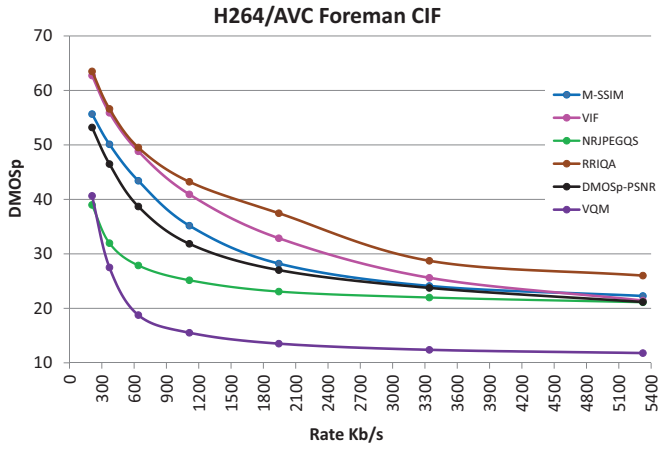


Figure 1.47: QAM comparison figures for Foreman CIF and H264/AVC codec in Intra mode

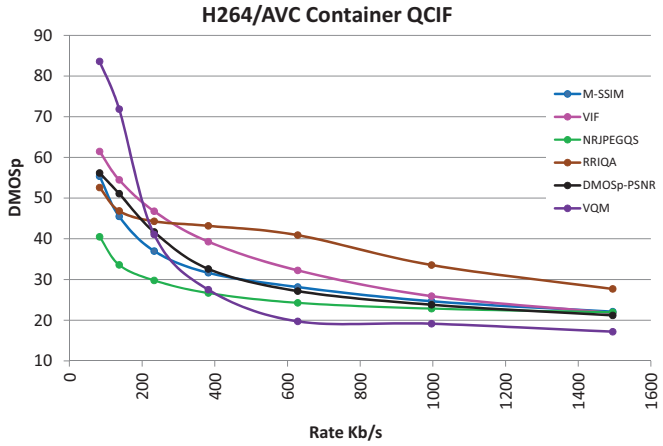


Figure 1.48: QAM comparison figures for Container QCIF and H264/AVC codec in Intra mode

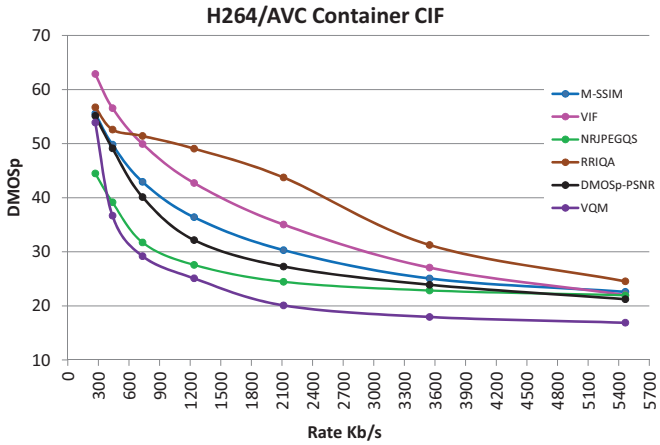


Figure 1.49: QAM comparison figures for Container QCIF and H264/AVC codec in Intra mode

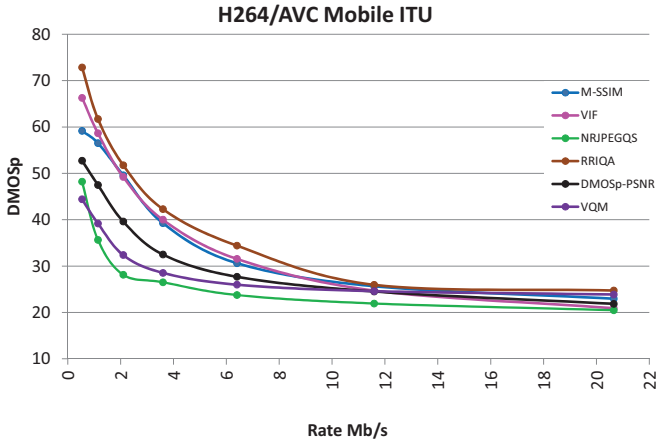


Figure 1.50: QAM comparison figures for Mobile ITU and H264/AVC codec in Intra mode

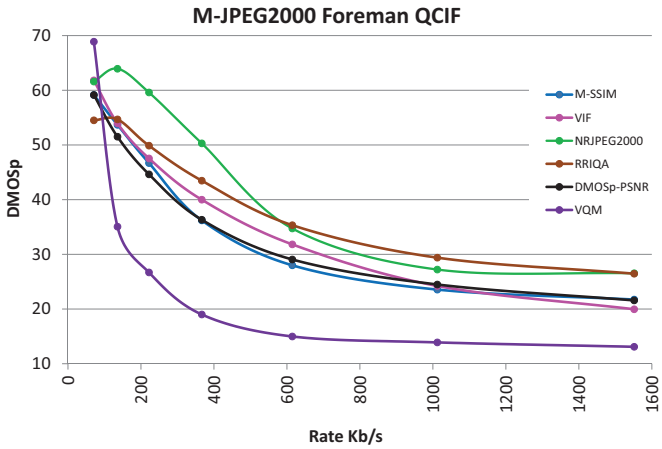


Figure 1.51: QAM comparison figures for Foreman QCIF and JPEG2000 codec

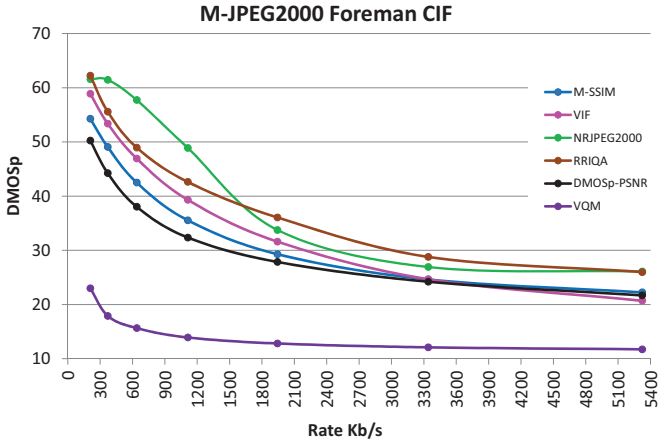


Figure 1.52: QAM comparison figures for Foreman CIF and JPEG2000 codec

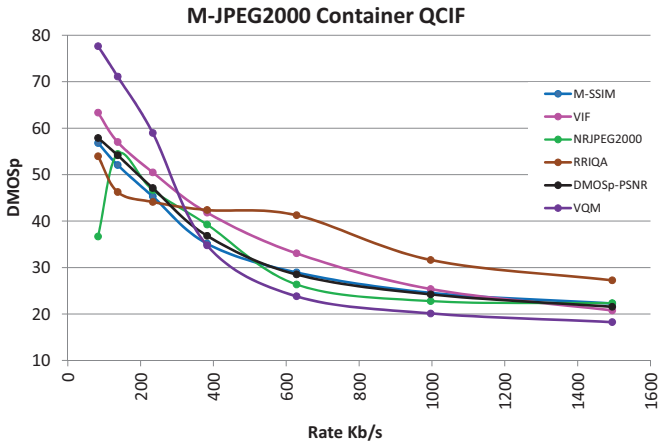


Figure 1.53: QAM comparison figures for Container QCIF and JPEG2000 codec

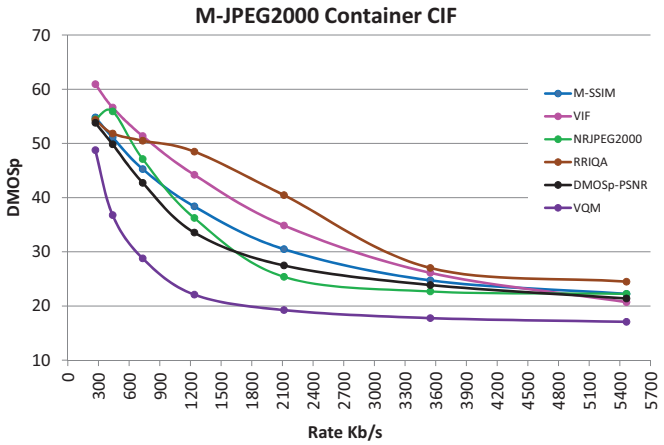


Figure 1.54: QAM comparison figures for Container CIF and JPEG2000 codec

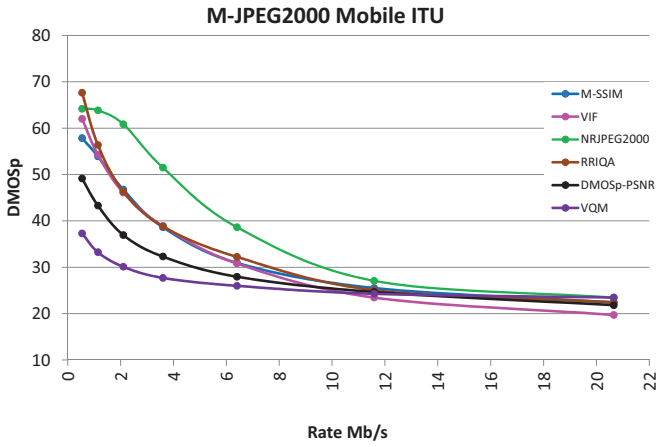


Figure 1.55: QAM comparison figures for Mobile ITU and JPEG2000 codec

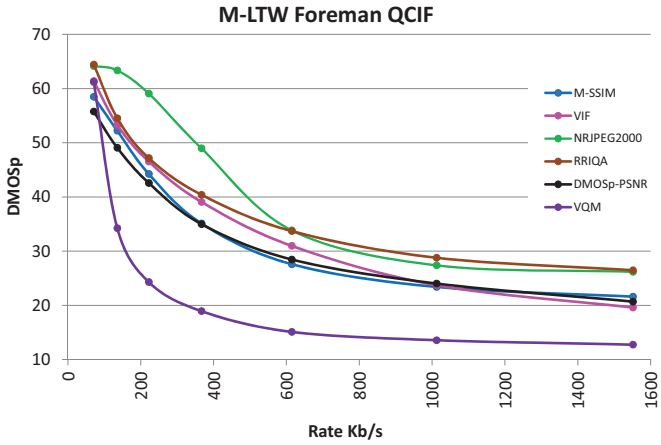


Figure 1.56: QAM comparison figures for Foreman QCIF and M-LTW codec

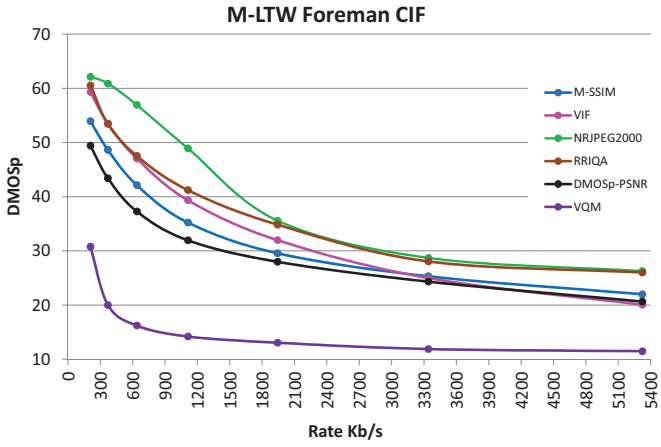


Figure 1.57: QAM comparison figures for Foreman CIF and M-LTW codec

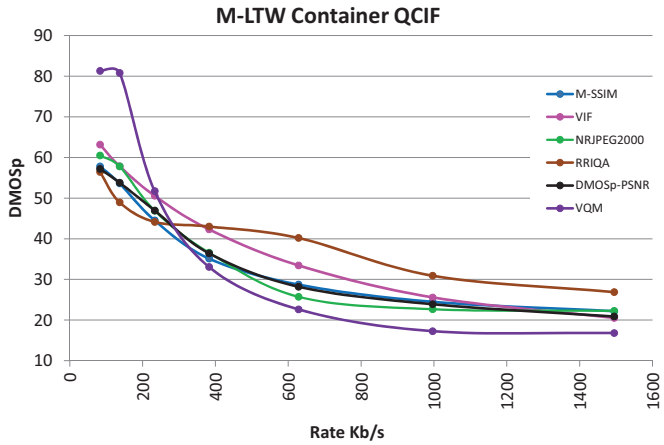


Figure 1.58: QAM comparison figures for Container QCIF and M-LTW codec

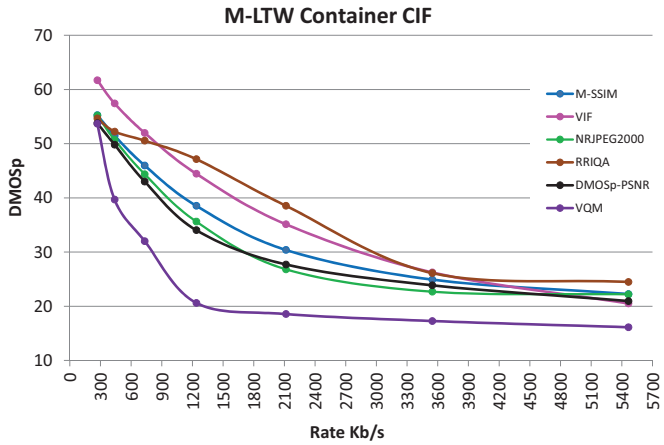


Figure 1.59: QAM comparison figures for Container CIF and M-LTW codec

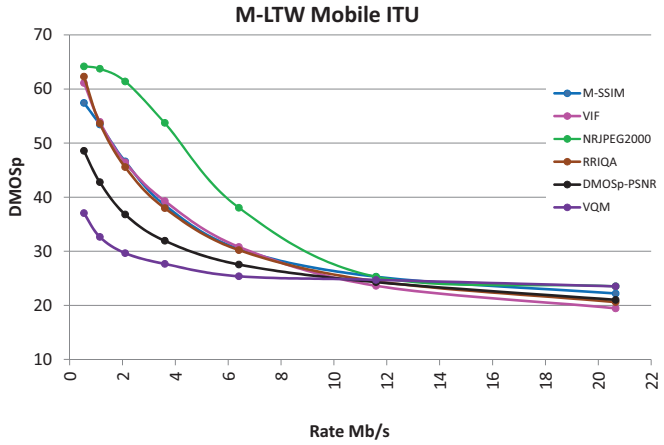


Figure 1.60: QAM comparison figures for Mobile ITU and M-LTW codec

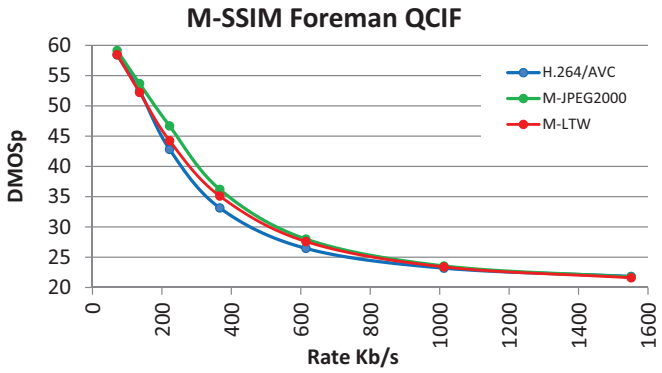


Figure 1.61: Encoders comparison figures for MSSIM - Foreman QCIF

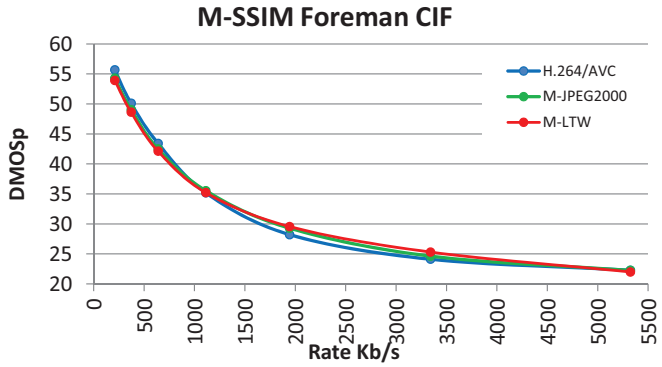


Figure 1.62: Encoders comparison figures for MSSIM - Foreman CIF

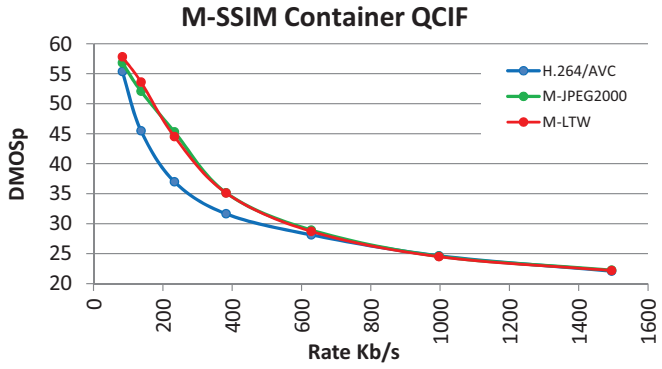


Figure 1.63: Encoders comparison figures for MSSIM - Container QCIF

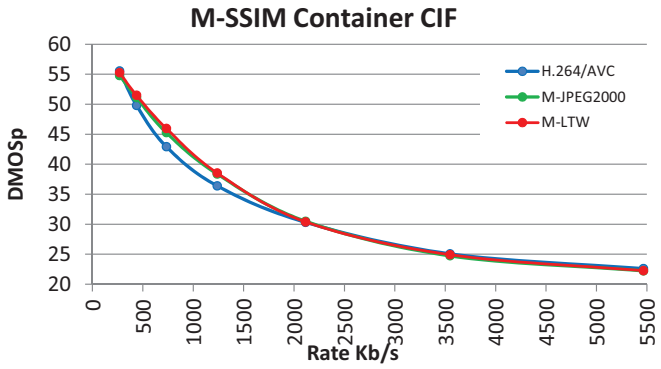


Figure 1.64: Encoders comparison figures for MSSIM - Container CIF

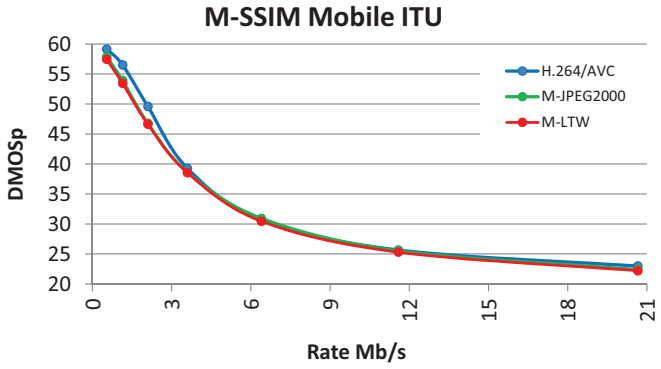


Figure 1.65: Encoders comparison figures for MSSIM - Mobile ITU

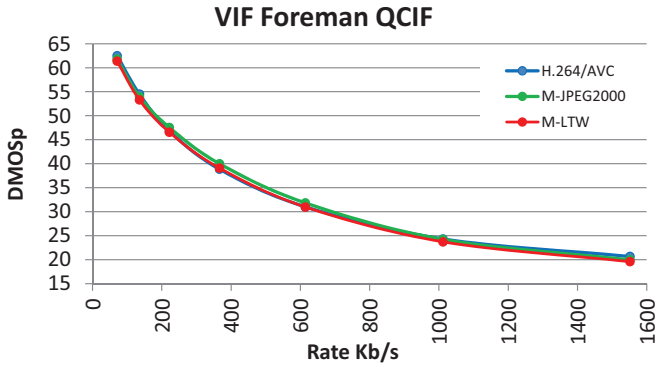


Figure 1.66: Encoders comparison figures for VIF - Foreman QCIF

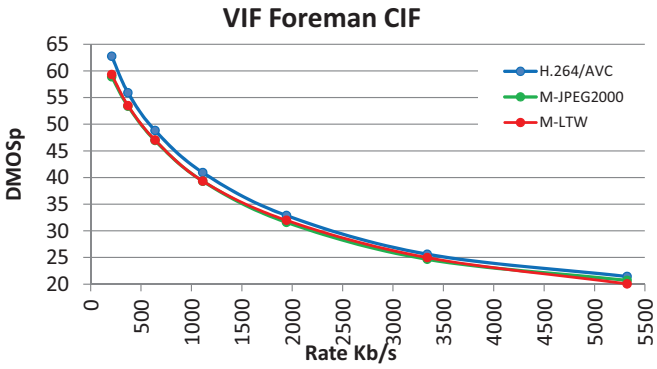


Figure 1.67: Encoders comparison figures for VIF - Foreman CIF

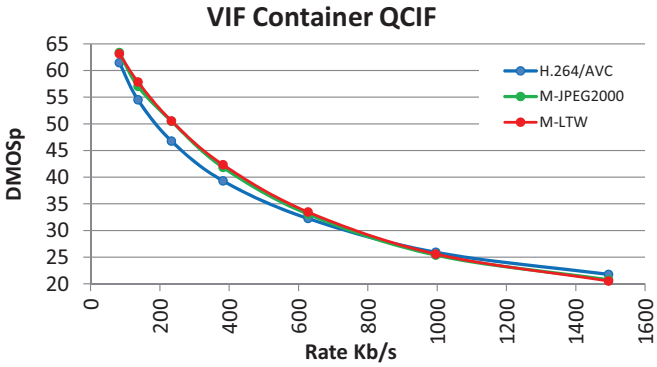


Figure 1.68: Encoders comparison figures for VIF - Container QCIF

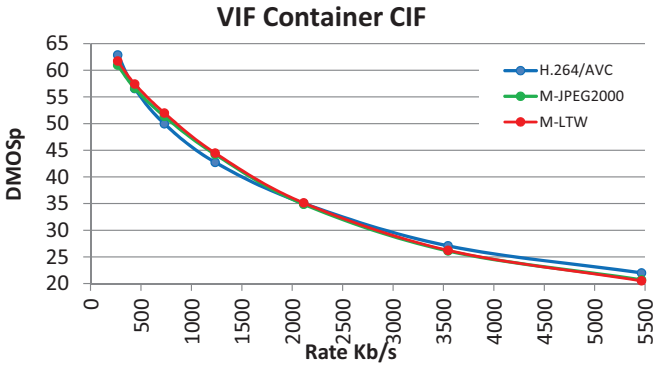


Figure 1.69: Encoders comparison figures for VIF - Container CIF

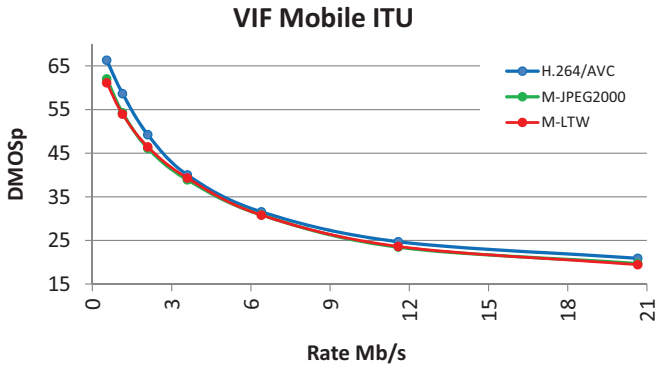


Figure 1.70: Encoders comparison figures for VIF - Mobile ITU

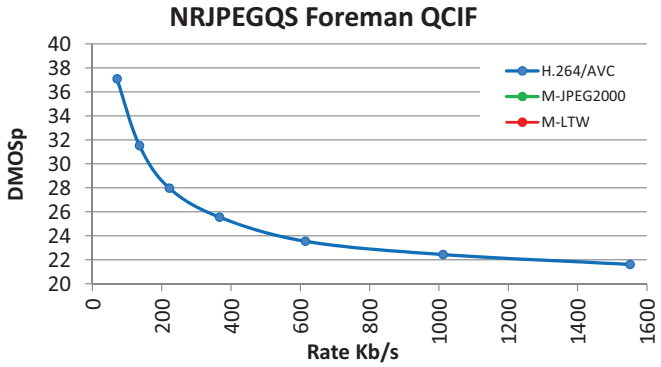


Figure 1.71: Encoders comparison figures for NRJPEGQS - Foreman QCIF

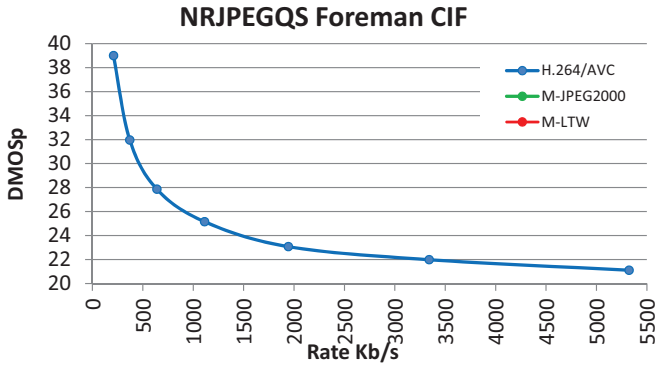


Figure 1.72: Encoders comparison figures for NRJPEGQS - Foreman CIF

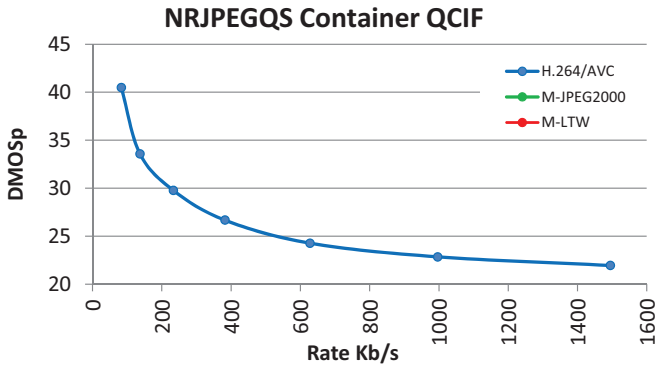


Figure 1.73: Encoders comparison figures for NRJPEGQS - Container QCIF

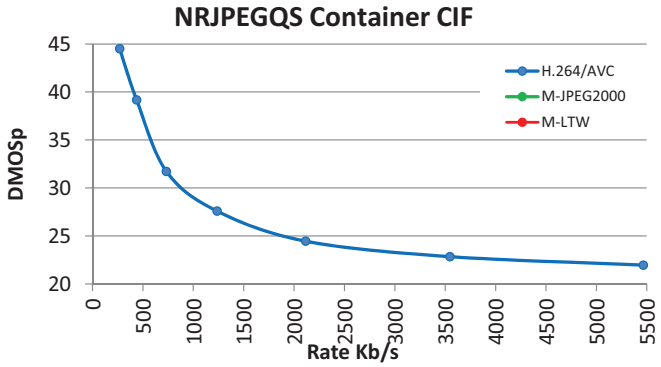


Figure 1.74: Encoders comparison figures for NRJPEGQS - Container CIF

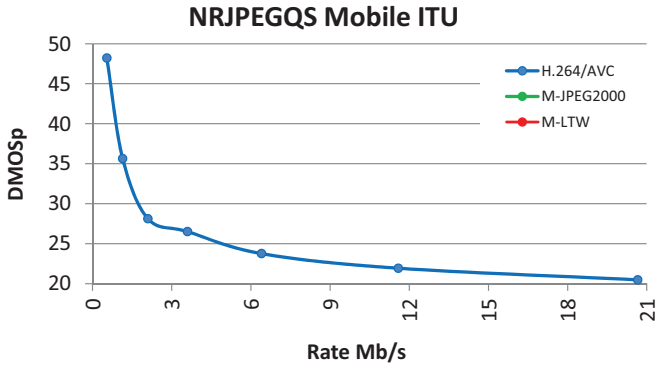


Figure 1.75: Encoders comparison figures for NRJPEGQS - Mobile ITU

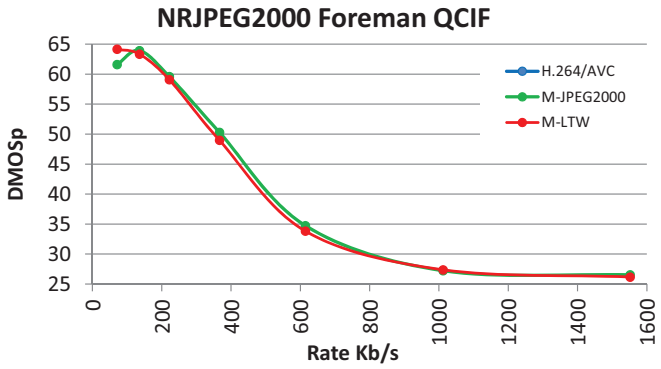


Figure 1.76: Encoders comparison figures for NRJPEG2000 - Foreman QCIF

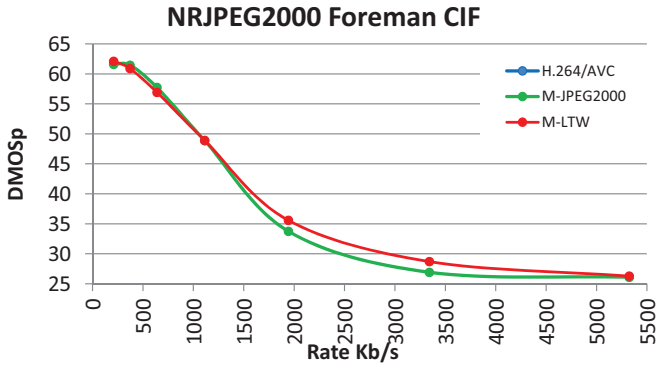


Figure 1.77: Encoders comparison figures for NRJPEG2000 - Foreman CIF

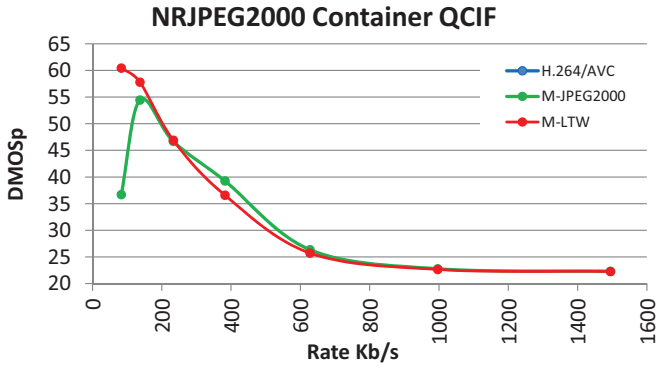


Figure 1.78: Encoders comparison figures for NRJPEG2000 - Container QCIF

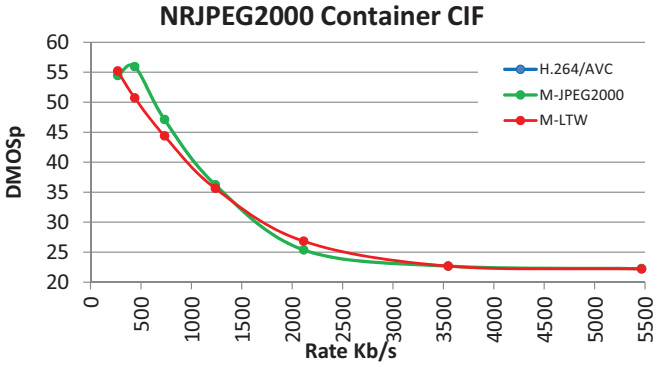


Figure 1.79: Encoders comparison figures for NRJPEG2000 - Container CIF

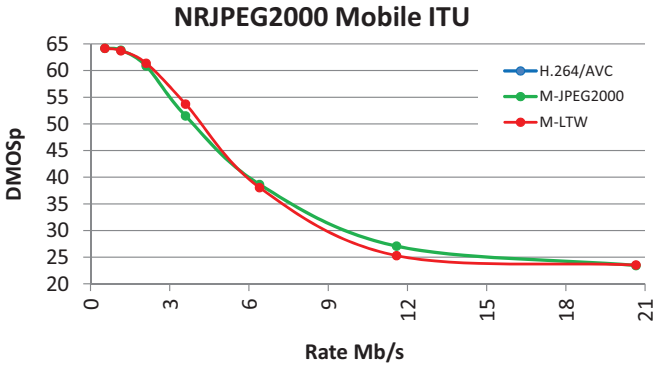


Figure 1.80: Encoders comparison figures for NRJPEG2000 - Mobile ITU

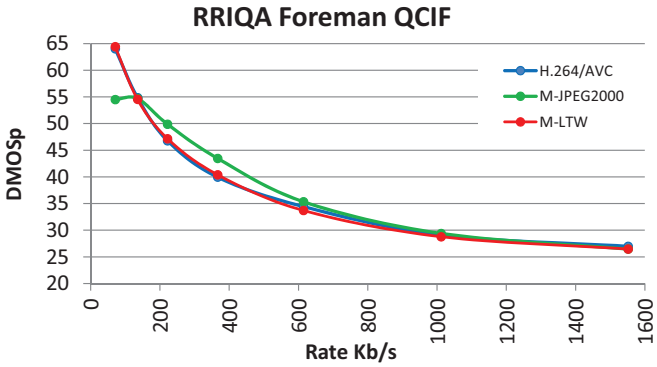


Figure 1.81: Encoders comparison figures for RRIQA - Foreman QCIF

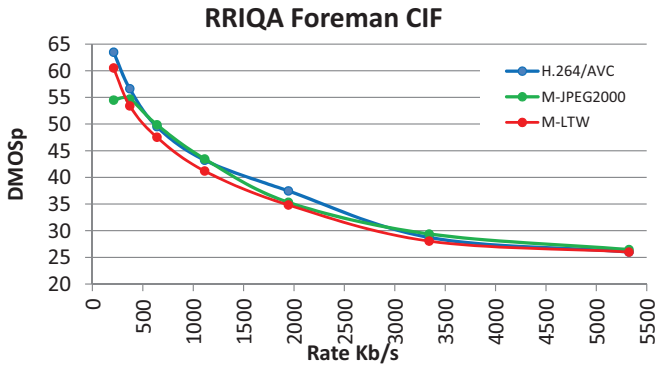


Figure 1.82: Encoders comparison figures for RRIQA - Foreman CIF

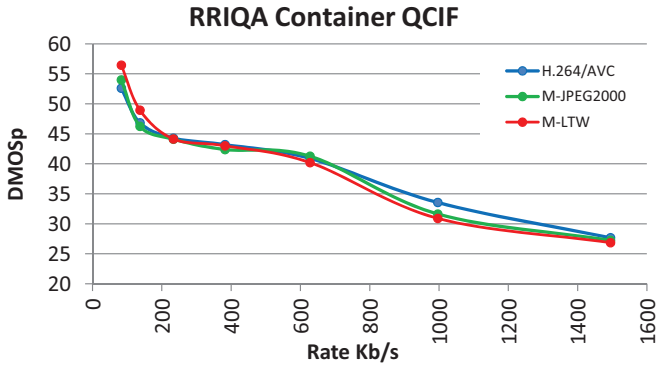


Figure 1.83: Encoders comparison figures for RRIQA - Container QCIF

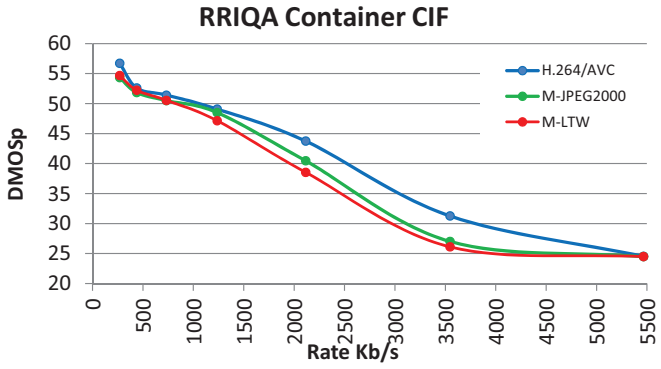


Figure 1.84: Encoders comparison figures for RRIQA - Container CIF

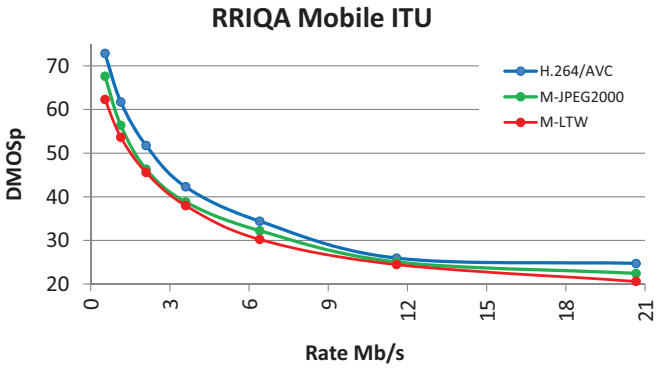


Figure 1.85: Encoders comparison figures for RRIQA - Mobile ITU

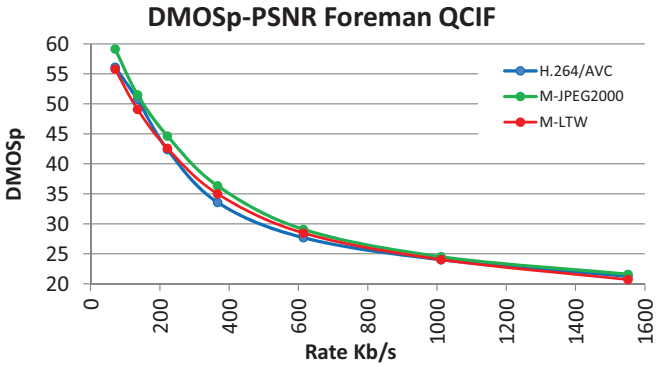


Figure 1.86: Encoders comparison figures for DMOSp-PSNR - Foreman QCIF

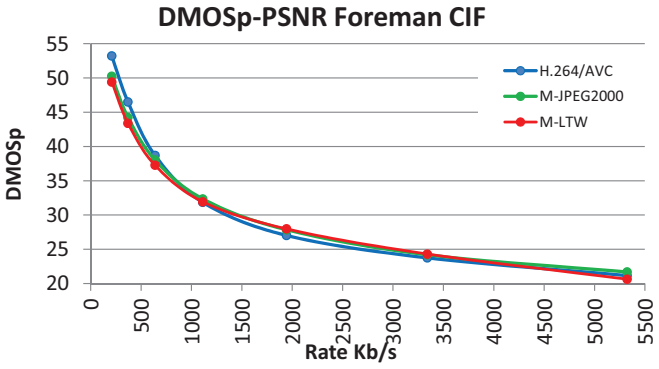


Figure 1.87: Encoders comparison figures for DMOSp-PSNR - Foreman CIF

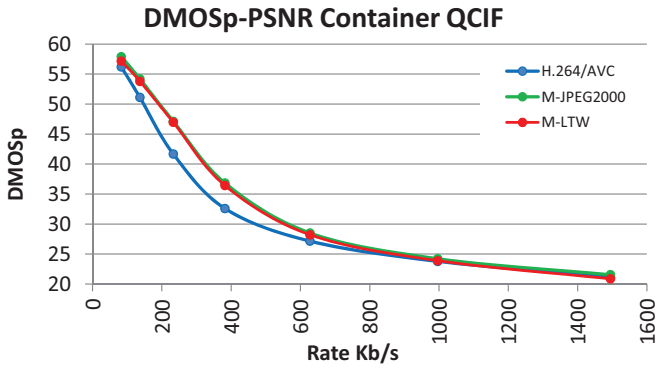


Figure 1.88: Encoders comparison figures for DMOSp-PSNR - Container QCIF

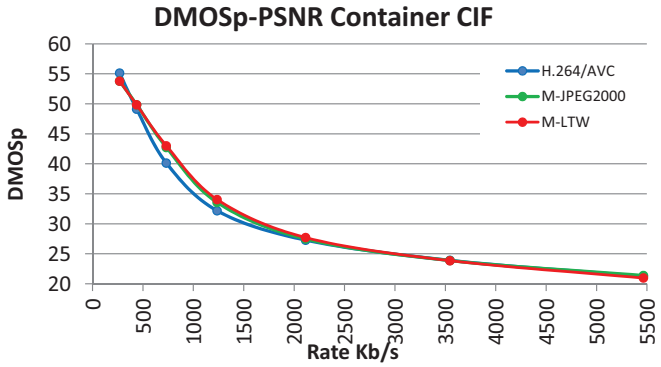


Figure 1.89: Encoders comparison figures for DMOSp-PSNR - Container CIF

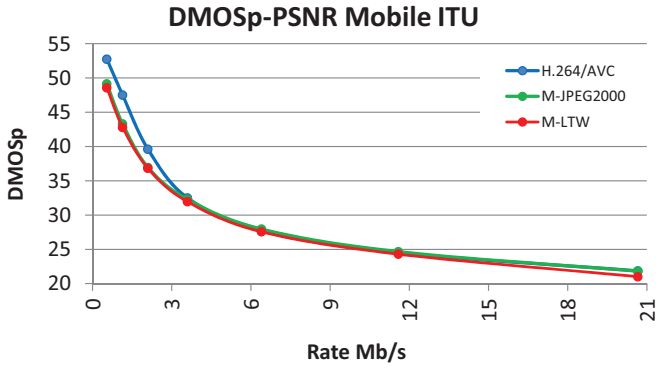


Figure 1.90: Encoders comparison figures for DMOSp-PSNR - Mobile ITU

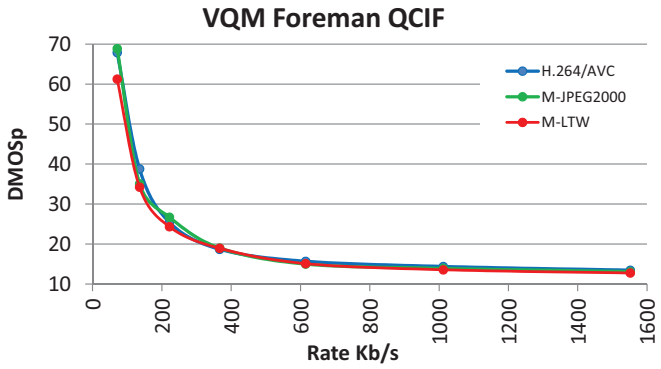


Figure 1.91: Encoders comparison figures for VQM - Foreman QCIF

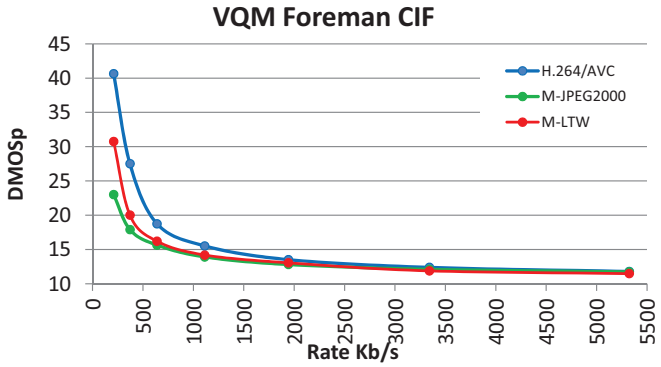


Figure 1.92: Encoders comparison figures for VQM - Foreman CIF

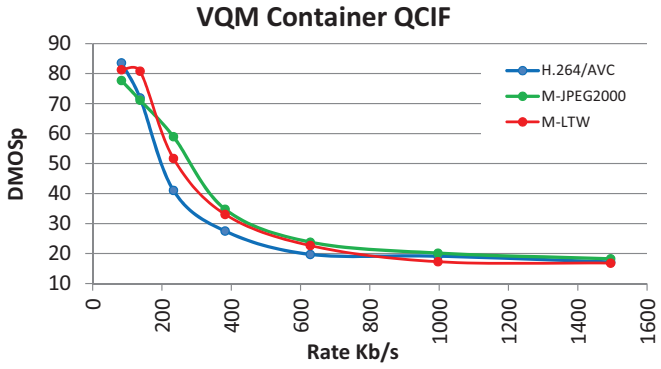


Figure 1.93: Encoders comparison figures for VQM - Container QCIF

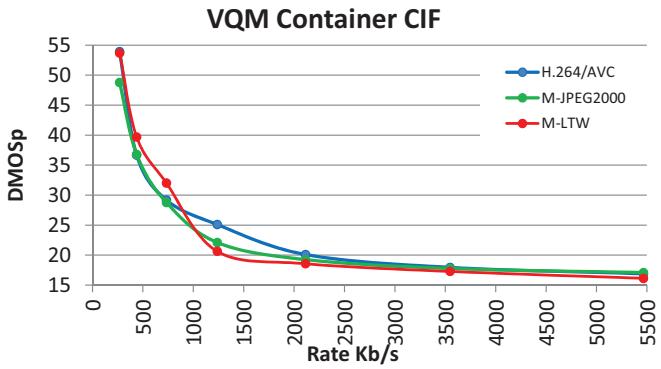


Figure 1.94: Encoders comparison figures for VQM - Container CIF

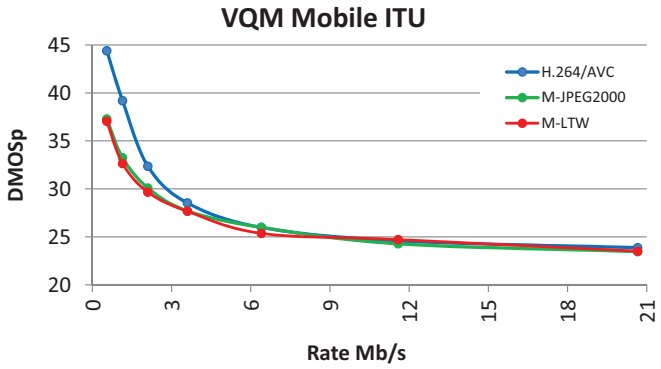


Figure 1.95: Encoders comparison figures for VQM - Mobile ITU

Appendix I

Acronyms

Bibliography

[1] Zhenghua Yu, Hong Ren Wu, S. Winkler, and Tao Chen. Vision-model-based impairment metric to evaluate blocking artifacts in digital video. *Proceedings of the IEEE*, 90(1):154–169, 2002.

[2] Zhou Wang, A.C. Bovik, and Ligang Lu. Why is image quality assessment so difficult? In *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, volume 4, pages IV–3313–IV–3316, 2002.

[3] ITU Telecomunication Standardization Sector of ITU. Itu-r bt.500-11 methodology for the subjective assessment of the quality of television pictures. Technical report, ITU, 2002.

[4] ITU Telecomunication Standardization Sector of ITU. Itu-r bt.500-12 methodology for the subjective assessment of the quality of television pictures - ihs, inc. Technical report, ITU, 2009.

[5] ITU Telecomunication Standardization Sector of ITU. Tu-t p.910 (04/2008) subjective video quality assessment methods for multimedia applications, 2008.

[6] Hamid Rahim Sheikh, Alan Conrad Bovik, and Gustavo de Veciana. An information fidelity criterion for image quality assessment using natural scene statistics. *IEEE Transactions on Image Processing*, 14(12), 2005.

[7] A.M. van Dijk and J.-B. Martens. Quality assessment of compressed images: A comparison between two methods. In *Image Processing, 1996. Proceedings., International Conference on*, volume 1, pages 25–28 vol.2, 1996.

[8] K.T. Tan, M. Ghanbari, and D.E. Pearson. An objective measurement tool for {MPEG} video quality. *Signal Processing*, 70(3):279 – 294, 1998.

[9] Margaret H. Pinson and Stephen Wolf. Comparing subjective video quality testing methodologies. *Proc. SPIE 5150, Visual Communications and Image Processing 2003*, pages 573–582, June 2003.

[10] H. R. Wu and K. R. Rao. *Digital Video Image Quality and Perceptual Coding (Signal Processing and Communications)*. CRC Press, Inc., Boca Raton, FL, USA, 2005.

[11] Philip Corriveau, Christina Gojmerac, Bronwen Hughes, and Lew Stelmach. All subjective scales are not created equal: The effects of context on different scales. *Signal Processing*, 77(1):1 – 9, 1999.

[12] Zhou Wang and A.C. 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.

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

[14] Zhou Wang and Alan C. Bovik. *Modern Image Quality Assessment*. Synthesis Lectures on Image, Video, and Multimedia Processing. Morgan & Claypool Publishers, 2006.

[15] K. Seshadrinathan, R. Soundararajan, A.C. Bovik, and L.K. Cormack. Study of subjective and objective quality assessment of video. *Image Processing, IEEE Transactions on*, 19(6):1427–1441, 2010.

[16] J. Korhonen and Junyong You. Peak signal-to-noise ratio revisited: Is simple beautiful? In *Quality of Multimedia Experience (QoMEX), 2012 Fourth International Workshop on*, pages 37–38, 2012.

[17] S. Winkler. Issues in vision modeling for perceptual video quality assessment. *Signal Processing*, 78(2), 1999.

[18] Z. Wang, H. R. Sheikh, and A. C. Bovik. *The Handbook of Video Databases: Design and Applications*, chapter 41 Objective Video Quality Assessment, pages 1041–1078. CRC Press, 2003.

[19] K. Brunnstrom, D. Hands, F. Speranza, and A. Webster. Vqeg validation and itu standardization of objective perceptual video quality metrics [standards in a nutshell]. *Signal Processing Magazine, IEEE*, 26(3):96–101, 2009.

[20] Farzad Ebrahimi, Matthieu Chamik, and Stefan Winkler. Jpeg vs. jpeg2000: An objective comparison of image encoding quality. In *Proceedings of SPIE Applications of Digital Image Processing*, page 300308, 2004.

[21] Michael Yuen and H.R. Wu. A survey of hybrid mc/dpcm/dct video coding distortions. *Signal Processing*, 70(3):247 – 278, 1998.

[22] R. Leung and D. Taubman. Minimizing the perceptual impact of visual distortion in scalable wavelet compressed video. In *Image Processing, 2006 IEEE International Conference on*, pages 633–636, 2006.

[23] Ying Luo and Rabab K. Ward. Removing the blocking artifacts of block-based dct compressed images. *IEEE Transactions on Image Processing*, 12(7), July 2003.

[24] R. Kakarala and R. Bagadi. A method for signalling block-adaptive quantization in baseline sequential jpeg. In *TENCON 2009 - 2009 IEEE Region 10 Conference*, pages 1–6, 2009.

[25] KinTak U., Nian Ji, Dongxu Qi, and Zesheng Tang. An adaptive quantization technique for jpeg based on non-uniform rectangular partition. In Ying Zhang, editor, *Future Wireless Networks and Information Systems*, volume 143 of *Lecture Notes in Electrical Engineering*, pages 179–187. Springer Berlin Heidelberg, 2012.

[26] Quoc Bao Do, M. Luong, and A. Beghdadi. A new perceptually adaptive method for deblocking and deringing. In *Information Science, Signal Processing and their Applications (ISSPA), 2012 11th International Conference on*, pages 533–538, 2012.

[27] W. Li, O. Egger, and M. Kunt. Efficient quantization noise reduction device for subband image coding schemes. In *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, volume 4, pages 2209–2212 vol.4, 1995.

[28] Andrew B. Watson, Gloria Y. Yang, Joshua A. Solomon, and John Villasenor. Visibility of wavelet quantization noise. *IEEE TRANSACTIONS ON IMAGE PROCESSING*, 6(8):1164–1175, 1997.

[29] Ngai-Fong Law, Wan-Chi Siu, and Degan Feng. Suppression of ringing artifacts with an adaptive shrinkage algorithm. In *Communications, Computers and Signal Processing, 1999 IEEE Pacific Rim Conference on*, pages 181–184, 1999.

[30] V.K. Nath and D. Hazarika. Blocking artifacts suppression in wavelet transform domain using local wiener filtering. In *Emerging Trends and Applications in Computer Science (NCETACS), 2012 3rd National Conference on*, pages 93–97, 2012.

[31] J. Oliver and M.P. Malumbres. Fast and efficient spatial scalable image compression using wavelet lowertrees. In *IEEE Data Compression Conference*, Snowbird, UT, 2003.

[32] Marcus J Nadenau, Stefan Winkler, David Alleysson, and Murat Kunt. Human vision models for perceptually optimized image processing–a review. *Proceedings of the IEEE*, page 32, 2000.

[33] Marcus Nadenau. *Integration of human color vision models into high quality image compression*. PhD thesis, STI, Lausanne, 2000.

[34] L. K. Cormack. *Handbook of Image and Video Processing*, chapter 4.1 Computational models, of early human vision, pages 271–287. Academic Press, 2000.

[35] G. Westheimer. *Handbook of Perception and Human Performance*, volume Vol.1 Chap. 4, chapter The eye as an optical instrument. John Wiley & Sons, 1986.

[36] David R. Williams. Topography of the foveal cone mosaic in the living human eye. *Vision Research*, 28(3):433 – 454, 1988.

[37] Jeffrey Lubin. Digital images and human vision. chapter The use of psychophysical data and models in the analysis of display system performance, pages 163–178. MIT Press, Cambridge, MA, USA, 1993.

[38] Sanghoon Lee, M.S. Pattichis, and A.C. Bovik. Foveated video quality assessment. *Multimedia, IEEE Transactions on*, 4(1):129–132, 2002.

[39] Zhou Wang and A.C. Bovik. Embedded foveation image coding. *Image Processing, IEEE Transactions on*, 10(10):1397–1410, 2001.

[40] Michael P. Eckert and Andrew P. Bradley. Perceptual quality metrics applied to still image compression. *Signal Processing*, 70(3):177 – 200, 1998.

[41] D.C. Hood and M.A. Finkelstein. *Sensitivity to light*, volume 1, chapter Handbook of Perception and Human Performance. 1986.

[42] Randolph Blake and Rober Sekuler. *Perception*, chapter Ch5. Spatial Vision and Form Perception, pages 151–192. 2005.

[43] F.W. Campbell and J.G. Robson. Application of fourier analysis to the visibility of gratings. *Jounal of Physiology*, 197:551–566, 1968.

[44] F.W. Campbell and C Blakemore. On the existence of neurones in the human visual system selectively sensitive to the orientation and size of retinal images. *Journal of Physiology (London)*, 203:237–260, 1969.

[45] Selig Hecht. The visual discrimination of intensity and the weber-fechner law. *Journal of General Psychology*, 7(2):235–267, 1924.

[46] Kathy T. Mullen. The contrast sensitivity of human colour vision to red-green and blue-yelow chromatic gratings. *Journal of Physiology*, pages 381–400, 1985.

[47] Stefan Winkler. *Digital video quality: vision models and metrics*. Wiley, 2005.

[48] S. Daly. Engineering observations from spatiovelocity and spatio-temporal visual models. In *Proc. SIPIE*, volume 3299, pages 180–191, San Jose, CA, 1998.

[49] Jan J. Koenderink and Andrea J. van Doorn. Spatiotemporal contrast detection threshold surface is bimodal. *Opt. Lett.*, 4(1):32–34, Jan 1979.

[50] J. G. ROBSON. Spatial and temporal contrast-sensitivity functions of the visual system. *J. Opt. Soc. Am.*, 56(8):1141–1142, Aug 1966.

[51] D. H. Kelly. Spatiotemporal variation of chromatic and achromatic contrast thresholds. *J. Opt. Soc. Am.*, 73(6):742–749, Jun 1983.

[52] Jian Yang and Walter Makous. Spatiotemporal separability in contrast sensitivity. *Vision Research*, 34(19):2569 – 2576, 1994.

[53] Dawei W. Dong. Spatiotemporal inseparability of natural images and visual sensitivities. In *In Computational, Neural & Ecological Constraints of Visual Motion Processing, J.M. Zanker & J. Zeil (Eds*, pages 371–380. Springer Verlag, 1999.

[54] M. A. Georgeson and G. D. Sullivan. Contrast constancy: Deblurring in human vision by spatial frequency channels. *Journal of Physiology*, 252(3):627–656, 1975.

[55] J. Fiser, P. J. Bex, and W. Makous. Contrast conservation in human vision. *Vision Research*, 43(25):2637–48, 2003.

[56] Michael A. Webster and Eriko Miyahara. Contrast adaptation and the spatial structure of natural images. *J. Opt. Soc. Am. A*, 14(9):2355–2366, Sep 1997.

[57] Damon M. Chandler and Sheila S. Hemami. Suprathreshold image compression based on contrast allocation and global precedence. In *in Proc. SPIE Human Vision and Electronic Imaging VIII*, pages 73–86, 2003.

[58] VQEG. Final report from the video quality experts group on the validation of objective models of video quality assessment. phase I, Marz 2000.

[59] VQEG. Final report from the video quality experts group on the validation of objective models of video quality assessment. phase II, August 2003.

[60] S. Winkler and P. Mohandas. The evolution of video quality measurement: From psnr to hybrid metrics. *Broadcasting, IEEE Transactions on*, 54(3):660–668, 2008.

[61] F. Porikli, A. Bovik, C. Plack, G. AlRegib, J. Farrell, P. Le Callet, Quan Huynh-Thu, S. Moller, and S. Winkler. Multimedia quality assessment [dsp forum]. *Signal Processing Magazine, IEEE*, 28(6):164–177, 2011.

[62] Stephen Wolf and Margaret H. Pinson. Spatial-temporal distortion metric for in-service quality monitoring of any digital video system, 1999.

[63] M. Masry, S. S. Hemami, and Y. Sermadevi. A scalable wavelet-based video distortion metric and applications. *IEEE Trans. Cir. and Sys. for Video Technol.*, 16(2):260–273, September 2006.

[64] ITU Telecomunication Standardization Sector of ITU. Itu-t rec j.144. objective perceptual video quality measurement techniques for digital cable television in the presence of a full reference, March 2001.

[65] Video Quality Experts Group (VQEG). Final report from the video quality experts group on the validation of objective models of multimedia quality assessment, phase i, March 2008.

[66] ITU Telecomunication Standardization Sector of ITU. Itu-t rec j.247. objective perceptual multimedia video quality measurement in the presence of a full reference, August 2008.

[67] ITU Telecomunication Standardization Sector of ITU. Itu-t rec j.246. perceptual audiovisual quality measurement techniques for multimedia services over digital cable television networks in the presence of a reduced bandwidth reference, August 2008.

[68] U. Engelke and H-J Zepernick. Perceptual-based quality metrics for image and video services: A survey. In *Next Generation Internet Networks, 3rd EuroNGI Conference on*, pages 190–197, 2007.

[69] S. Winkler. Video quality measurement standards current status and trends. In *Information, Communications and Signal Processing, 2009. ICICS 2009. 7th International Conference on*, pages 1–5, 2009.

[70] S. Chikkerur, V. Sundaram, M. Reisslein, and L.J. Karam. Objective video quality assessment methods: A classification, review, and performance comparison. *Broadcasting, IEEE Transactions on*, 57(2):165–182, 2011.

[71] Weisi Lin and C.-C. Jay Kuo. Perceptual visual quality metrics: A survey. *Journal of Visual Communication and Image Representation*, 22(4):297 – 312, 2011.

[72] P.C. Teo and D.J. Heeger. Perceptual image distortion. In *Image Processing, 1994. Proceedings. ICIP-94., IEEE International Conference*, volume 2, pages 982–986 vol.2, 1994.

[73] Christian J. van den Branden Lambrecht and Olivier Verscheure. Perceptual quality measure using a spatiotemporal model of the human visual system. In *Storage and Retrieval for Image and Video Databases*, volume 2668, pages 450–461, 1996.

[74] Andrew B. Watson, James Hu, and John F Mcgowan Iii. Dvq: A digital video quality metric based on human vision. *Journal of Electronic Imaging*, 10:20–29, 2001.

[75] J. Malo, A.M. Pons, and J.M. Artigas. Subjective image fidelity metric based on bit allocation of the human visual system in the {DCT} domain. *Image and Vision Computing*, 15(7):535 – 548, 1997.

[76] Andrew B Watson. Toward a perceptual video-quality metric. In *Photonics West'98 Electronic Imaging*, pages 139–147. International Society for Optics and Photonics, 1998.

[77] Zhou Wang, Alan C. Bovik, Ligang Lu, and Jack L. Kouloheris. Foveated wavelet image quality index, 2001.

[78] A. Cavallaro and S. Winkler. Segmentation-driven perceptual quality metrics. In *Image Processing, 2004. ICIP '04. 2004 International Conference on*, volume 5, pages 3543–3546 Vol. 5, 2004.

[79] Stefan Winkler. Perceptual distortion metric for digital color video, 1999.

[80] Stefan Winkler. Quality metric design: a closer look, 2000.

[81] Eli Peli. Contrast in complex images. *J. Opt. Soc. Am. A*, 7(10):2032–2040, Oct 1990.

[82] C.J. Van Den Branden Lambrecht. A working spatio-temporal model of the human visual system for image restoration and quality assessment applications. In *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, volume 4, pages 2291–2294 vol. 4, 1996.

[83] Andrew B. Watson. The cortex transform: Rapid computation of simulated neural images. *Computer Vision, Graphics, and Image Processing*, 39(3):311 – 327, 1987.

[84] Scott Daly. Digital images and human vision. chapter The visible differences predictor: an algorithm for the assessment of image fidelity, pages 179–206. MIT Press, Cambridge, MA, USA, 1993.

[85] E. P. Simoncelli, W. T. Freeman, E. H. Adelson, and D. J. Heeger. Shiftable multiscale transforms. *IEEE Trans. Inf. Theor.*, 38(2):587–607, September 2006.

[86] Andrew B Watson. Visual optimization of dct quantization matrices for individual images. In *Proc. AIAA Computing in Aerospace*, volume 9, pages 286–291, 1993.

[87] Marcus J. Nadenau, Julien Reichel, and Murat Kunt. Performance comparison of masking models based on a new psychovisual test method with natural scenery stimuli. *Signal Processing: Image Communication*, 17(10):807 – 823, 2002.

[88] Andrew B Watson and Joshua A Solomon. Model of visual contrast gain control and pattern masking. *JOSA A*, 14(9):2379–2391, 1997.

[89] A.B. Watson. Perceptual optimization of dct color quantization matrices. In *Image Processing, 1994. Proceedings. ICIP-94., IEEE International Conference*, volume 1, pages 100–104 vol.1, 1994.

[90] Stefan Winkler and Ruth Campos. Video quality evaluation for internet streaming applications, 2003.

[91] Y. Sermadevi and S.S. Hemami. Linear programming optimization for video coding under multiple constraints. In *Data Compression Conference, 2003. Proceedings. DCC 2003*, pages 53–62, 2003.

[92] Arthur A Webster, Coleen T Jones, Margaret H Pinson, Stephen D Voran, and Stephen Wolf. Objective video quality assessment system based on human perception. In *IS&T/SPIE's Symposium on Electronic Imaging: Science and Technology*, pages 15–26. International Society for Optics and Photonics, 1993.

[93] Stephen Wolf and Margaret Pinson. Technical report tr-02-392 - video quality measurement techniques. Technical report, National Telecommunications & Information Administration, 2002.

[94] M.H. Pinson and S. Wolf. A new standardized method for objectively measuring video quality. *Broadcasting, IEEE Transactions on*, 50(3):312–322, 2004.

[95] Stephen Wolf and Margaret H Pinson. Low bandwidth reduced reference video quality monitoring system. In *First Int'l Workshop on Video Proc. and Quality Metrics*, 2005.

[96] Zhou Wang, A.C. Bovik, and B.L. Evan. Blind measurement of blocking artifacts in images. In *Image Processing, 2000. Proceedings. 2000 International Conference on*, volume 3, pages 981–984 vol.3, 2000.

[97] H.R. Wu and M. Yuen. A generalized block-edge impairment metric for video coding. *Signal Processing Letters, IEEE*, 4(11):317–320, 1997.

[98] A.C. Bovik and Shizhong Liu. Dct-domain blind measurement of blocking artifacts in dct-coded images. In *Acoustics, Speech, and Signal Processing, 2001. Proceedings. (ICASSP '01). 2001 IEEE International Conference on*, volume 3, pages 1725–1728 vol.3, 2001.

[99] Shizhong Liu and A.C. Bovik. Efficient dct-domain blind measurement and reduction of blocking artifacts. *Circuits and Systems for Video Technology, IEEE Transactions on*, 12(12):1139–1149, 2002.

[100] S.A. Karunasekera and N.G. Kingsbury. A distortion measure for blocking artifacts in images based on human visual sensitivity. *Image Processing, IEEE Transactions on*, 4(6):713–724, 1995.

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

[102] P. Marziliano, F. Dufaux, S. Winkler, and T. Ebrahimi. A no-reference perceptual blur metric. In *Image Processing. 2002. Proceedings. 2002 International Conference on*, volume 3, pages III–57–III–60 vol.3, 2002.

[103] Pina Marziliano, Frederic Dufaux, Stefan Winkler, and Touradj Ebrahimi. Perceptual blur and ringing metrics: application to jpeg2000. *Signal Processing: Image Communication*, 19(2):163–172, 2004.

[104] 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.

[105] S. Saha and R. Vemuri. Effect of image activity on lossy and lossless coding performance. In *Data Compression Conference, 2000. Proceedings. DCC 2000*, pages 570–, 2000.

[106] Paolo Gastaldo, Rodolfo Zunino, Ingrid Heynderickx, and Elena Vicario. Objective quality assessment of displayed images by using neural networks. *Signal Processing: Image Communication*, 20(7):643–661, 2005.

[107] Marco Montenovo, Alessandro Perot, Marco Carli, Paolo Cicchetti, and Alessandro Neri. Objective quality evaluation of video services. In *Second International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, 2006.

[108] S. Winkler, E. D. Gelasca, and T. Ebrahimi. Perceptual quality assessment for video watermarking. In *International Conference on Information Technology: Coding and Computing, 2002. Proceedings.*, pages 90–94, April 2002.

[109] Zhou Wang, Guixing Wu, Hamid R. Sheikh Member, Eero P. Simoncelli Senior Member, En hui Yang, Senior Member, and Alan C. Bovik. Quality-aware images. *IEEE Transactions on Image Processing*, 15:1680–1689, 2006.

[110] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *Image Processing, IEEE Transactions on*, 13(4):600–612, 2004.

[111] Andrew B. Watson and Lindsay Kreslake. Measurement of visual impairment scales for digital video, 2001.

[112] Marcia G. Ramos and Sheila S. Hemami. Suprathreshold wavelet coefficient quantization in complex stimuli: psychophysical evaluation and analysis. *J. Opt. Soc. Am. A*, 18(10):2385–2397, Oct 2001.

[113] Damon M. Chandler and Sheila S. Hemami. Additivity models for suprathreshold distortion in quantized wavelet-coded images, 2002.

[114] A. Srivastava, A. B. Lee, E. P. Simoncelli, and S.-C. Zhu. On advances in statistical modeling of natural images. *J. Math. Imaging Vis.*, 18(1):17–33, January 2003.

[115] Zhou Wang and A.C. Bovik. A universal image quality index. *Signal Processing Letters, IEEE*, 9(3):81–84, 2002.

[116] Zhou Wang, Alan C. Bovik, and Eero P. Simoncelli. *Handbook of Image and Video Processing*, chapter 8.3 Structural Approaches to Image Quality Assessment. Aca, 2005.

[117] Z. Wang, L. Lu, and A. Bovik. Video quality assessment using structural distortion measurement. In *Proceedings IEEE International Conference of Image Processing*, volume 3, pages 65–68, September 2002.

[118] Zhou Wang, Ligang Lu, and Alan C. Bovik. Video quality assessment based on structural distortion measurement. *Signal Processing: Image Communication, special issue on "Objective Video Quality Metrics"*, 19(2):121–132, February 2004.

[119] Z. Wang, E.P. Simoncelli, and A.C. Bovik. Multiscale structural similarity for image quality assessment. In *Signals, Systems and Computers, 2004. Conference Record of the Thirty-Seventh Asilomar Conference on*, volume 2, pages 1398–1402 Vol.2, 2003.

[120] Zhou Wang and E.P. Simoncelli. Translation insensitive image similarity in complex wavelet domain. In *Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on*, volume 2, pages 573–576, 2005.

[121] Zhou Wang and Eero P. Simoncelli. An adaptative linear system framework for image distortion analysis. In *Proc. 12th IEEE Intl. Conf. Image Processing Vol III, pp 1160-1163, Sep 2005.*, 2005.

[122] Eero P Simoncelli. Modeling the joint statistics of images in the wavelet domain. In *SPIE's International Symposium on Optical Science, Engineering, and Instrumentation*, pages 188–195. International Society for Optics and Photonics, 1999.

[123] Zhou Wang and Eero P. Simoncelli. Reduced-reference image quality assessment using a wavelet-domain natural image statistic model. In *Human Vision and Electronic Imaging X, Proc. SPIE, vol. 5666.*, 2005.

[124] H.R. Sheikh, A.C. Bovik, and L. Cormack. No-reference quality assessment using natural scene statistics: Jpeg2000. *Image Processing, IEEE Transactions on*, 14(11):1918–1927, 2005.

[125] E.P. Simoncelli. Statistical models for images: compression, restoration and synthesis. In *Signals, Systems amp; Computers, 1997. Conference Record of the Thirty-First Asilomar Conference on*, volume 1, pages 673–678 vol.1, 1997.

[126] R.W. Buccigrossi and E.P. Simoncelli. Image compression via joint statistical characterization in the wavelet domain. *Image Processing, IEEE Transactions on*, 8(12):1688–1701, 1999.

[127] Martin J Wainwright, Eero P Simoncelli, and Alan S Willsky. Random cascades on wavelet trees and their use in analyzing and modeling natural images. *Applied and Computational Harmonic Analysis*, 11(1):89–123, 2001.

[128] J. Korhonen, N. Burini, Junyong You, and E. Nadernejad. How to evaluate objective video quality metrics reliably. In *Quality of Multimedia Experience (QoMEX), 2012 Fourth International Workshop on*, pages 57–62, 2012.

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

[130] 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.

[131] Eric C. Larson and Damon M. Chandler. Most apparent distortion: full-reference image quality assessment and the role of strategy. *Journal of Electronic Imaging*, 19(1):011006, 2010.

[132] Patrick Le Callet and Florent Autrusseau. Subjective quality assessment irccyn/ivc database, 2005. http://www.irccyn.ec-nantes.fr/ivcdb/.

[133] Media Information and Communications Technology Laboratory. Toyama image database. OnLine - http://160.26.142.130/mictdb.html, 2010.

[134] D.M. Chandler and S.S. Hemami. Vsnr: A wavelet-based visual signal-to-noise ratio for natural images. *Image Processing, IEEE Transactions on*, 16(9):2284–2298, Sept 2007.

[135] Nikolay Ponomarenko, Federica Battisti, Karen Egiazarian, Jaakko Astola, and Vladimir Lukin. Metrics performance comparison for color image database. In *Fourth international workshop on video processing and quality metrics for consumer electronics*, volume 27, 2009.

[136] U. Engelke, H.J. Zepernick, and M. Kusuma. Wireless imaging quality database. OnLine - http://www.bth.se/tek/rcg.nsf/pages/wiq-db, 2010.

[137] P. V. Vu and D. M. Chandler. Vis3: An algorithm for video quality assessment via analysis of spatial and spatiotemporal slices. *Journal of Electronic Imaging (JEI)*, 23(1), 2014.

[138] Technische Universität München, Institute for Data Processing. TUM LDV Multi Format Test Set, 2011.

[139] Video Quality Experts Group (VQEG). Vqeg fr-tv phase i database. http://www.its.bldrdoc.gov/vqeg/downloads.aspx.

[140] Video Quality Experts Group (VQEG). Vqeg hdtv phase i database. http://www.its.bldrdoc.gov/vqeg/downloads.aspx.

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

[142] Ann Marie Rohaly, Philip Corriveau, John Libert, Arthur Webster, Vittorio Baroncini, John Beerends, and Jean-Louis Blin. Video quality experts group: Current results and future directions, 2000.

[143] ISO/IEC 14496-10:2003. Coding of audiovisual objects part 10: advanced videocoding. ITUT Recommendation H264 Advanced video codingfor generic audiovisual services, 2003.

[144] ISO/IEC 15444-1. Jpeg 2000 image coding system. part 1:core coding system,, 2000.

[145] J. Oliver and M.P. Malumbres. Low-complexity multiresolution image compression using wavelet lower trees. *IEEE Transactions on Circuits and Systems for Video Technology*, 17(11):1437–1444, Nov 2006.

[146] Carlos T. Calafate, P. Manzoni, and Manuel P. Malumbres. Speeding up the evaluation of multimedia streaming applications in MANETs using

HMMs. In *Proceedings of the 7th ACM international symposium on Modeling, analysis and simulation of wireless and mobile systems*, pages 315–322, 2004.

[147] IEEE. IEEE 802.11 WG. 802.11e Specific requirements Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications: Amendment 8: Medium Access Control (MAC) Quality of Service Enhancements, 2005.

[148] Carlos T. Calafate, Manuel P. Malumbres, and P. Manzoni. Performance of H.264 compressed video streams over 802.11b based MANETs. In *Proceedings of the 24th International Conference on Distributed Computing Systems Workshops - W7: EC (ICDCSW'04) - Volume 7*, pages 776 – 781, 2004.