16

Video Quality Metrics

Mylène C. Q. Farias
Department of Computer Science
University of Brasília (UnB)
Brazil

1. Introduction

Digital video communication has evolved into an important field in the past few years. There have been significant advances in compression and transmission techniques, which have made possible to deliver high quality video to the end user. In particular, the advent of new technologies has allowed the creation of many new telecommunication services (e.g., direct broadcast satellite, digital television, high definition TV, video teleconferencing, Internet video).
To quantify the performance of a digital video communication system, it is important to have a measure of video quality changes at each of the communication system stages. Since in the majority of these applications the transformed or processed video is destined for human consumption, humans will ultimately decide if the operation was successful or not. Therefore, human perception should be taken into account when trying to establish the degree to which a video can be compressed, deciding if the video transmission was successful, or deciding whether visual enhancements have provided an actual benefit.
Measuring the quality of a video implies a direct or indirect comparison of the test video with the original video. The most accurate way to determine the quality of a video is by measuring it using psychophysical experiments with human subjects (ITU-R, 1998). Unfortunately, psychophysical experiments are very expensive, time-consuming and hard to incorporate into a design process or an automatic quality of service control. Therefore, the ability to measure video quality accurately and efficiently, without using human observers, is highly desirable in practical applications. Good video quality metrics can be employed to monitor video quality, compare the performance of video processing systems and algorithms, and to optimize the algorithms and parameter settings for a video processing system.
With this in mind, fast algorithms that give a physical measure (objective metric) of the video quality are used to obtain an estimate of the quality of a video when being transmitted, received or displayed. Customarily, quality measurements have been largely limited to a few objective measures, such as the mean absolute error (MAE), the mean square error (MSE), and the peak signal-to-noise ratio (PSNR), supplemented by limited subjective evaluation. Although the use of such metrics is fairly standard in published literature, it suffers from one major weakness. The outputs of these measures do not always correspond well with human judgements of quality.
In the past few years, a big effort in the scientific community has been devoted to the development of better video quality metrics that correlate well with the human perception of quality (Daly, 1993; Lubin, 1993; Watson et al., 2001; Wolf et al., 1991). Although much

Source: Digital Video, Book edited by: Floriano De Rango,
ISBN 978-953-7619-70-1, pp. 500, February 2010, INTECH, Croatia, downloaded from SCIYO.COM

has been done in the last ten years, there are still a lot of challenges to be solved since most of the achievements have been in the development of full-reference video quality metrics that evaluate compression artifacts. Much remains to be done, for example, in the area of no-reference and reduced-reference quality metrics. Also, given the growing popularity of video delivery services over IP networks (e.g. internet streaming and IPTV) or wireless channel (e.g. mobile TV), there is a great need for metrics that estimate the quality of the video in these applications.
In this chapter, we introduce several aspects of video quality. We give a brief description of the Human Visual System (HVS), discuss its anatomy and a number of phenomena of visual perception that are of particular relevance to video quality. We also describe the main characteristics of modern digital video systems, focusing on how visible errors (artifacts) are perceived in digital videos. The chapter gives a description of a representative set of video quality metrics. We also discuss recent developments in the area of video quality, including the work of the Video Quality Experts Group (VQEG).

2. The Human Visual System (HVS)

In the past century, the knowledge about the human visual system (HVS) has increased tremendously. Although much more needs to be learned before we can claim to understand it, the current state of the art of visual information-processing mechanisms is sufficient to provide important information that can be used in the design of video quality metrics. In fact, results in the literature show that video quality metrics that use models based on the characteristics of the HVS have better performance, i.e., give predictions that are better correlated with the values given by human observers (VQEG, 2003).
In this section, we introduce basic aspects of the anatomy and psychophysical features of the HVS that are considered relevant to video processing algorithms and, more specifically, to the design of video quality metrics.

2.1 Anatomy of the HVS

The eyes are far more than a simple camera. A more accurate description would be a self-focusing, self-adjusting for light intensity, and self-cleaning camera that provides a real-time output to a very advanced computer. The main components of the eye are the cornea, the pupil, the lens, and the fluids that fill the eye. A transverse section of the human eye is shown in Fig. 1.
The *optics* of the eye is composed by three major elements: the cornea, the pupil, and the lens. The light (visual stimulus) comes in through the optics and it is projected on the retina – the membrane located on the back of the eye. The optics works just like camera lens and their function is to project a clear and focused image on the retina – the retinal image. Given the physical limitation of the optics, the retinal image is only an approximation of the original image (the visual stimulus). As a result, the retinal image main contain some distortions, among which the most noticeable one is blurring. Since the response of optics is roughly linear, shift-invariant, and low-pass, the resulting retinal image can be approximated by convolving the input visual image with a blurring point spread function (PSF) (Marr, 1982).
The *retina* has the main function of translating the incoming light into nerve signals that can be understood by the brain. It has the shape of a plate and it is composed of many layers of neurons, as depicted in Fig. 2. The light projected on the retina has to pass through several layers before it reaches the photoreceptors cells and is absorbed by the pigment layer.

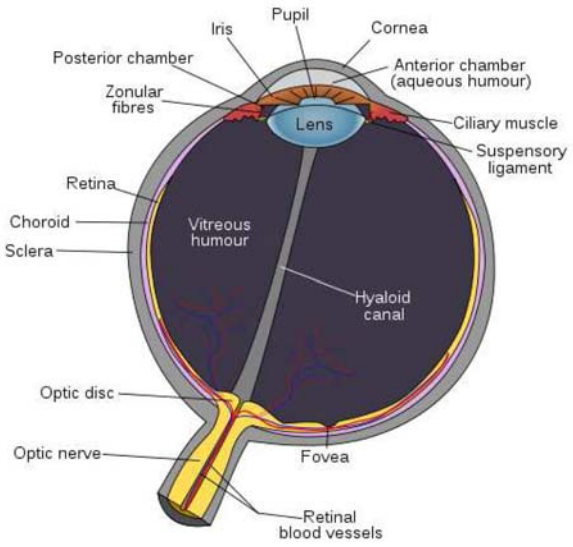


Fig. 1. Transverse section of the human eye (Wikimedia Commons, 2007).

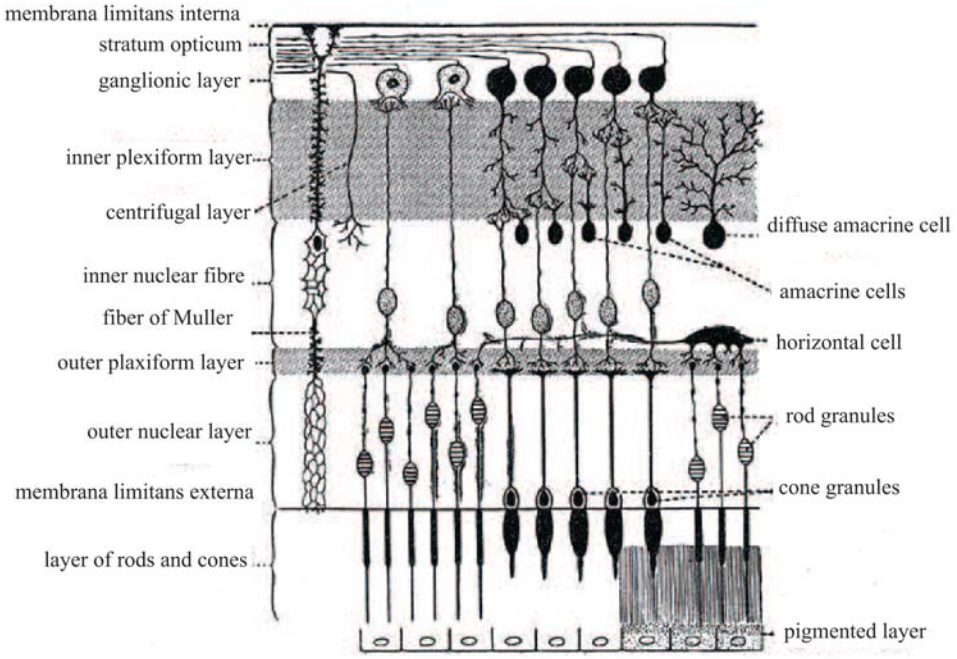


Fig. 2. Plan of retinal neurons. The retina is a stack of several neuronal layers. Light has to pass these layers (from top to bottom) to hit the photoreceptors (layer of rods and cones). The signal propagates through the bipolar and horizontal cells (middle layers) and, then, to the amacrine and ganglion cells. (Adapted from H. Grey (Grey, 1918))

The photoreceptor cells are specialized neurons that convert light energy into signals which can then be understood by the brain. There are two types of photoreceptors cells: *cones* and *rods*. Observe from Fig. 2 that the names are inspired by the shape of the cells. The rods are responsible for vision in low-light conditions. Cones are responsible for vision in normal high-light conditions, color vision, and have the ability to see fine details.

There are three types of cones, which are classified according to the spectral sensitivity of their photochemicals. The tree types are known as *L-cones*, *M-cones*, and *S-cones*, which stand for long, medium, and short wavelengths cones, respectively. Each of them has peak sensitivities around 570nm, 540nm, and 440nm, respectively. These differences are what makes color perception possible. The incoming light from the retina is split among the three types of cones, according to its spectral content. This generates three visual streams that roughly correspond to the three primary colors red, green, and blue.

There are roughly 5 million cones and 100 million rods in a human eye. But their distribution varies largely across the surface of the retina. The center of the retina has the highest density of cones and ganglion cells (neurons that carry the electrical signal from the eye to the brain through the optic nerve). This central area is called *fovea* and is only about half a millimeter in diameter. As we move away from it, the density of both cones and ganglion cells falls off rapidly. Therefore, the fovea is responsible for our fine-detail vision and, as a consequence, we cannot perceive the entire visual stimulus at uniform resolution.

The majority of cones in the retina are L- and M-cones, with S-cones accounting for less than 10% of the total number of cones. Rods, on the other hand, dominate outside the fovea. As a consequence, it is much easier to see dim objects when they are located in the peripheral field of vision. Looking at Fig. 1, we can see that there is a hole or *blind spot*, where the optic nerve is. In this region there are no photoreceptors.

The signal collected from the photoreceptors has to pass through several layers of neurons in the retina (retinal neurons) before being carried off to the brain by the optic nerve. As depicted in Fig. 2, different types of neurons can be found in the retina:

- *Horizontal cells* link receptors and bipolar cells by relatively long connections that run parallel to the retinal layers.
- *Bipolar cells* receive input from the receptors, many of them feeding directly into the retinal ganglion cells.
- *Amacrine cells* link bipolar cells and retinal ganglion cells.
- *Ganglion cells* collect information from bipolar and amacrine cells. Their axons form the optic nerve that leaves the eye through the optic disc and carries the output signal from the retina to other processing centers in the brain.

The signal leaves the eye through the *optic nerve*, formed by the axons of the ganglion cells. A scheme showing central connections of the optic nerves to the brain is depicted in Fig. 3. Observe that the optic nerves from the left and right eye meet at the *optic chiasm*, where the fibers are rearranged. About half of these fibers cross to the opposite side of the brain and the other half stay on the same side. In fact, the corresponding halves of the field of view (right and left) are sent to the left and right halves of the brain. Considering that the retinal images are reversed by the optics of the eye, the right side of the brain processes the left half (of the field of view) of both eyes, while the left side processes the right half of both eyes. This is illustrated by the red and blue lines in Fig. 3.

From the optic chiasm, the fibers are taken to several parts of the brain. Around 90% of them finish at the two *lateral geniculate body*. Besides serving as a relay station for signals from the

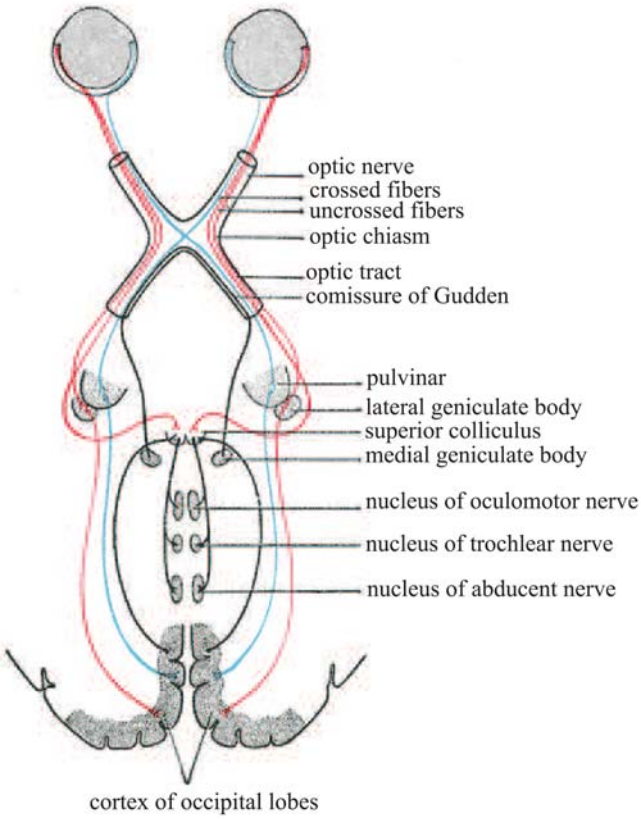


Fig. 3. Scheme showing central connections of the optic nerves and optic tracts. (Adapted from H. Grey (Grey, 1918))

retina to the visual cortex, the lateral geniculate body controls how much information is allowed to pass. From there, the fibers are taken to the visual cortex.

The *virtual cortex* is the region of the brain responsible for processing the visual information. It is located on the back of the cerebral hemispheres. The region that receives the information from the lateral geniculate body is called the *primary visual cortex* (also known as V1). In addition to V1, more than 20 other areas receiving visual input have been discovered, but little is known about their functionalities.

V1 is a region specialized on processing information about static and moving objects and recognizing patterns. There is a big variety of cells in V1 that have selective sensitivity to certain types of information. In other words, one particular cell may respond strongly to patterns of a certain orientation or to motion in a certain direction. Others are tuned to particular frequencies, color, velocities, etc. An interesting characteristic of these neurons is the fact that their outputs saturates as the input contrast increases.

The selectivity of the neurons in V1 is the heart of the multichannel organization characteristic of the human vision system. In fact, the neurons in V1 can be modeled as an octave-band Gabor filter bank, where the spatial frequency spectrum (in polar

representation) is sampled at octave intervals in the radial frequency dimension and at uniform intervals in the orientation dimension (Marr, 1982). This model is used by several algorithms in image processing and video quality assessment.

2.2 Perceptual features

A number of visual perception phenomena are a consequence of the characteristics of the optics of the human eye. The phenomena described in this section are of particular interest to the area of image processing and, more specifically to video quality.

2.2.1 Foveal and peripheral vision

The densities of the photoreceptors and ganglion cells in the retina are not uniform, increasing towards the center of the retina (fovea) and decreasing on the contrary direction. As a consequence, the resolution of objects in the visual field is also not uniform. The point where the observer fixates is projected on the fovea and, consequently, resolved with the highest resolution. The objects in the peripheral area are resolved with progressively lower resolution (peripheral vision).

2.2.2 Light adaptation

In the real world, the amount of light intensity varies tremendously, from dim (night) to high intensity (sun day). The HVS adapts to this large range by controlling the amount of light that enters the eye. This is done by increasing/decreasing the diameter of the pupils and, at the same time, adjusting the gain of post-receptor neurons in the retina. As a result, instead of coding absolute light intensities, the retina encodes the contrast of the visual stimulus.

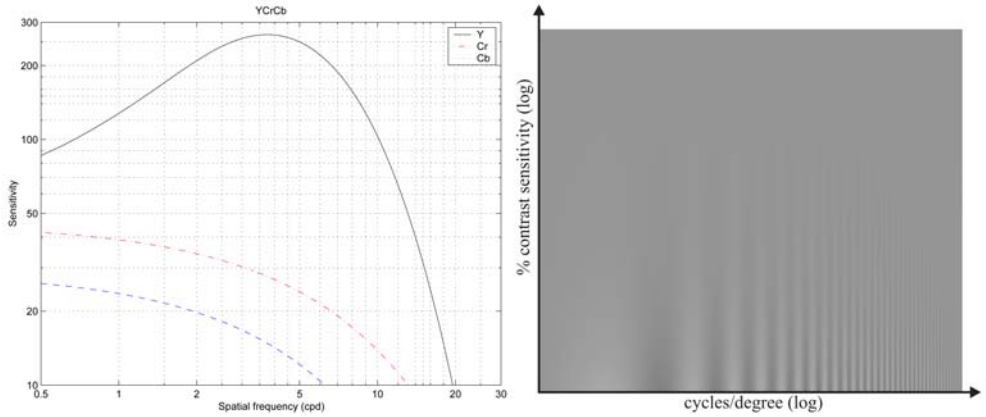
The phenomenon that keeps the contrast sensitivity over a wide range of light intensity is known as Weber's law:

$$\Delta I / I = K$$

where I is the background luminance, ΔI is the just noticeable incremental luminance over the background, and K is a constant called the Weber fraction.

2.2.3 Contrast Sensitivity Functions (CSF)

CSF models the sensitivity of the HVS as a function of the spatial frequency of the visual stimuli. A typical CSF is shown in Fig. 4(a). Spatial contrast sensitivity peaks at 3 cycles per degree (cpd), and declines more rapidly at higher than at lower spatial frequencies. Frequencies higher than 40 cpd (8 cpd scotopic) are undetectable even at maximum contrast. For illustration purposes, consider the image in Fig. 4(b) that corresponds to the intensities of a sinusoidal luminance grating. In this image, the spatial frequency (number of luminance cycles the grating repeats in one degree of visual angle) increases from left to right, while contrast (difference between the maximum and minimum luminance) increases from top to bottom. The shape of the visible lower part of the image gives an indication of our relative sensitivity to different spatial frequencies. If the perception of contrast were determined solely by the image contrast, then the alternating bright and dark bars should appear to have equal height across any horizontal line across the image. However, the bars are observed to be significantly higher at the middle of the image, following the shape of the CSF (see Fig. 4(a)).



(a) (b)

Fig. 4. (a) Contrast sensitivity functions for the three channels YCbCr (after Moore, 2002 (Moore, 2002)). (b) Pelli-Robson Chart, where spatial frequency increases from left to right, while contrast increases from top to bottom.

2.2.4 Masking and facilitation

Masking and facilitation are important aspects of the HVS in modeling the interactions between different image components present at the same spatial location. Specifically, these two effects refer to the fact that the presence of one image component (*the mask*) will decrease/ increase the visibility of another image component (*test signal*). The mask generally reduces the visibility of the test signal in comparison with the case where the mask is absent. However, the mask may sometimes facilitate detection as well. Usually, the masking effect is the strongest when the mask and the test signal have similar frequency content and orientations. Most quality metrics incorporate a model for masking and/or facilitation.

2.2.5 Pooling

Pooling refers to the task of arriving at a single measurement of quality from the outputs of the visual streams. It is not quite understood how the HVS performs pooling. But, it is clear that a perceptible distortion may be more annoying in some areas of the scene (such as human faces) than in others. Most quality metrics use the *Minkowski metric* to pool the error signals from the streams with different frequency and orientation selective and arrive at a fidelity measurement (de Ridder, 1992; 2001). The Minkowski metric is also used to combine information across spatial and temporal coordinates.

3. Digital video systems

In this section, we give a brief overview of the available video compression and transmission techniques and their impact on the quality of a digital video.

3.1 Video compression

Video compression (or video coding) is the process of converting a video signal into a format that takes up less storage space or transmission bandwidth. Given the video

transmission and storage requirements (up to 270 Mbits/s for Standard Definition and 1.5 Gbit/s for High Definition), video compression is an essential technology for applications such as digital television (terrestrial, cable or satellite transmission), optical storage/reproduction, mobile TV, videoconferencing and Internet video streaming (Poynton, 2003).

There are two types of compression: *lossy* and *lossless* compression (Bosi & Goldberg, 2002). Lossless compression algorithms have the characteristic of assuring perfect reconstruction of the original data. Unfortunately, this type of compression only allows around 2:1 compression ratios, which is not sufficient for video applications. Lossy compression is the type of compression most commonly used for video because it provides much bigger compression ratios. There is, of course, a trade-off: the higher the compression ratio, the lower the quality of the compressed video.

Compression is achieved by removing the redundant information from the video. There are four main types of redundancies that are typically explored by compression algorithms:

- *Perceptual redundancy*: Information of the video that cannot be easily perceived by the human observer and, therefore, can be discarded without significantly altering the quality of the video.
- *Temporal redundancy*: Pixels in successive video frames have great similarity. So, even though motion tend to change the position of blocks of pixels, it does not change their values and therefore their correlation.
- *Spatial redundancy*: There is a significant correlation among pixels around the same neighborhood in a frame.
- *Statistical redundancy*: This type of redundancy is related to the statistical relationship within the video data (bits and bytes).

Each stage of a video compression algorithm is responsible for mainly reducing one type of redundancy. Fig. 5 depicts the functional components in a typical video compression algorithm. Different algorithms differ in what tools are used in each stage. But, most of them share the same principles: motion compensation and block-based transform with subsequent quantization. Currently, there are several standards for video compression, which standardize the decoding process. The encoding process is not fixed, what leaves room for innovation.

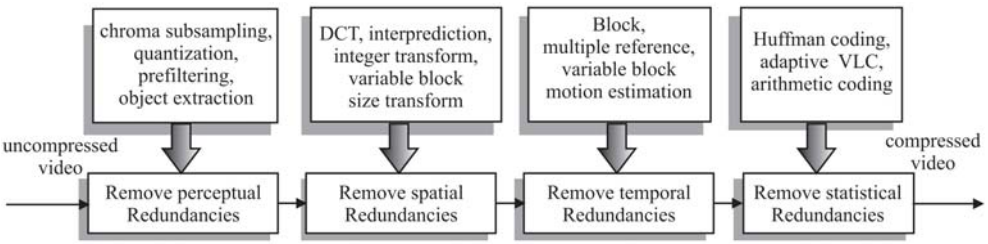


Fig. 5. Functional components in a typical video compression algorithm.

The most popular compression standards were produced by the Motion Picture Experts Group (MPEG) (ITU, 1998) and the Video Coding Experts (VCEG). The MPEG is a working group of the International Organization for Standardization (ISO) and of the International Electrotechnical Commission (IEC), formally known as ISO/IEC – JTC1/SC29/WG11. Among the standards developed by MPEG areMPEG-1,MPEG-2, andMPEG-4. The MPEG-2

is a very popular standard used not only for broadcasting, but also in DVDs (Haskell et al., 1997; ITU, 1998). The main advantage of MPEG-2 is its low cost, given its popularity and the large scale of production. MPEG-2 is also undoubtedly a very mature technology.
The VCEG is a working group of the Telecommunication Standardization Sector of the International Telecommunication Union (ITU-T). Among the standards developed by VCEG are the H.261 and H.263. A joint collaboration between MPEG and VCEG resulted in the development of the H.264, also known as MPEG-4 Part 10 or AVC (Advanced Video Coding) (Richardson, 2003; ITU, 2003). The H.264 represents a major advance in the technology of video compression, providing a considerable reduction of bitrate when compared to previous standards (Lambert et al., Jan. 2006). For the same quality level, H.264 provides a bitrate of about half the bitrate provided by MPEG-2.

3.2 Digital video transmission

Compressed video streams are mainly intended for transmission over communication networks. But, there are different types of video communication and streaming applications. Each one has particular operating conditions and properties. The channels used for video communication may be static or dynamic, packet-switched or circuit-switched. Also, the channels may support a constant or variable bit rate transmission, and may support some form of Quality of Service (QoS) or may only provide best effort support. Finally, the transmission may be point-to-point, multicast, and broadcast.
In most cases, after the video has been digitally compressed, the resulting bitstream is segmented into fixed or variable packets and multiplexed with other data types, such as audio. The next stage is the channel encoder, which will add error protection to the data. The characteristics of the specific video communication application will, of course, have a great impact on the quality of the video displayed at the receiver.

3.3 Common artifacts in digital video systems

An impairment is a property of the video that is perceived as undesirable, whether it is in the original or not. Impairments can be introduced during capture, transmission, storage, and/or display, as well as by any image processing algorithm (e.g. compression) that may be applied along the way (Yuen & Wu, 1998). They can be very complex in their physical descriptions and also in their perceptual descriptions. Most of them have more than one perceptual feature, but it is possible to have impairments that are relatively pure. To differentiate impairments from their perceptual features, we will use the term *artifact* to refer to the perceptual features of impairments and *artifact signal* to refer to the physical signal that produces the artifact.
The most common artifacts present in digital video are:

- *Blockiness* or *blocking* – A type of artifact characterized by a block pattern visible in the picture. It is due to the independent quantization of individual blocks (usually of 8x8 pixels in size) in block-based DCT coding schemes, leading to discontinuities at the boundaries of adjacent blocks. The blocking effect is often the most visible artifact in a compressed video, given its periodicity and the extent of the pattern. More modern codecs, like the H.264, use a deblocking filter to reduce the annoyance caused by this artifact.

- *Blur* or *blurring* – It is characterized for a loss of spatial detail and a reduction of edge sharpness. In the in the compression stage, blurring is introduced by the suppression of the high-frequency coefficients by coarse quantization.
- *Color bleeding* – It is characterized by the smearing of colors between areas of strongly differing chrominance. It results from the suppression of high-frequency coefficients of the chroma components. Due to chroma subsampling, color bleeding extends over an entire macroblock.
- *DCT basis image effect* – It is characterized by the prominence of a single DCT coefficient in a block. At coarse quantization levels, this results in an emphasis of the dominant basis image and reduction of all other basis images.
- *Staircase effect* – These artifacts occurs as a consequence of the fact that DCT basis are best suited for the representation of horizontal and vertical lines. The representation of lines with other orientations require higher-frequency DCT coefficients for accurate reconstruction. Therefore, when higher frequencies are lost, slanted lines appear.
- *Ringing* – Associated with the Gibbs phenomenon. It is more evident along high contrast edges in otherwise smooth areas. It is a direct result of quantization leading to high-frequency irregularities in the reconstruction. Ringing occurs with both luminance and chroma components.
- *Mosquito noise* – Temporal artifact that is seen mainly in smoothly textured regions as luminance/chrominance fluctuations around high contrast edges or moving objects. It is a consequence of the coding differences for the same area of a scene in consecutive frames of a sequence.
- *Flickering* – It occurs when a scene has a high texture content. Texture blocks are compressed with varying quantization factors over time, which results in a visible flickering effect.
- *Packet loss* – It occurs when parts of the video are lost in the digital transmission. As a consequence, parts (blocks) of video are missing for several frames.
- *Jitter* – It is the result of skipping regularly video frames to reduce the amount of video information that the system is required to encode or transmit. This creates motion perceived as a series of distinct snapshots, rather than smooth and continuous motion.

The performance of a particular digital video system can be improved if the type of artifact that is affecting the quality of the video is known (Klein, 1993). This type of information can also be used to enhance the video by reducing or eliminating the identified artifacts (Caviedes & Jung, 2001). In summary, this knowledge makes it possible to implement a complete system for detecting, estimating and correcting artifacts in video sequences. Unfortunately, there is not yet a good understanding of how visible/annoying these artifacts are, how the content influences their visibility/annoyance, and how they combine to produce the overall annoyance. A comprehensive *subjective* study of the most common types of artifacts is still needed.

An effort in this direction has been done by Farias *et al* (Farias, Moore, Foley & Mitra, 2002; Farias et al., 2003a;b; Farias, Foley & Mitra, 2004; Farias, Moore, Foley & Mitra, 2004). Their approach makes use of synthetic artifacts that look like "real" artifacts, yet are simpler, purer, and easier to describe. This approach makes it possible to control the type, proportion, and strength of the artifacts being tested and allows to evaluate the performance of different combination models of the artifact metrics. The results gathered from the psychophysical experiments performed by Farias *et al* show that the synthetic artifacts,

besides being visually similar to the real impairments, have similar visibility and annoyance properties. Their results also show that there is an interaction between among different types of artifacts. For example, the presence of noisy artifact signals seem to decrease the perceived strength of the other artifacts, while the presence of blurry artifact signals seem to increase it. The authors also modeled annoyance by combining the artifact perceptual strengths (MSV) using both a Minkowski metric and a linear model (de Ridder, 1992).

4. Subjective video quality assessment

Subjective experiments (also called psychophysical experiments) represent the most accurate way of measuring the quality of a video. In subjective experiments, a number of subjects (observers or participants) are asked to watch a set of test sequences and give judgements about their quality or the annoyance of the impairments. The average of the values collected for each test sequence are known as Mean Observer Score (MOS).

In general, subjective experiments are expensive and time-consuming. The design, execution, and data analysis consume a great amount of the experimenter's time. Running an experiment requires the availability of subjects, equipment, and physical space. As a result, the number of experiments that can be conducted is limited and, therefore, an appropriate methodology should be used to get the most out of the resources.

The International Telecommunication Union (ITU) has recommendations for subjective testing procedures. The two most important documents are the ITU-R Rec. BT.500-11 (ITU-R, 1998), targeted at television applications, and the ITU-T Rec. P.910 (ITU-T, 1999), targeted at multimedia applications. These documents give information regarding the standard viewing conditions, the criteria for selections of observers and test material, assessment procedures, and data analysis methods. Before choosing which method to use, the experimenter should take into account the application in mind and the accuracy objectives.

According to ITU, there are two classes of subjective assessments:

- *Quality assessments* – The judgements given by subjects are in a quality scale, i.e., how good or bad is the quality of the displayed video. These assessments establish the performance of systems under optimum conditions;
- *Impairment assessments* – The judgements given by subjects are in an impairment scale, i.e., how visible or imperceptible are the impairments in the displayed video. These assessments establish the ability of systems to retain quality under non-optimum conditions that relate to transmission.

According to the type of scale, quality or impairment judgements can be classified as *continuous* or *discrete*. Judgements can also be categorical or non-categorical, adjectival or numerical. Depending on the form of presentation of the stimulus (sequences), the assessment method can be classified as *double* or *single* stimulus. In the single stimulus approach the test sequence is presented by itself, while in the double stimulus method a pair of sequences (test sequence and the corresponding reference) are presented together.

The most popular assessment procedures of ITU-R Rec. BT.500-11 are:

- *Double Stimulus Continuous Quality Scale* (DSCQS) – This method is specially useful when the test conditions exhibit the full range of quality. The observer is shown multiple pairs of sequences consisting of a test sequence and the corresponding reference. The sequences have a short duration of around 10s and are presented twice, alternated by each other. The observers are not told which is the reference and which is the test sequence. In each trial, their positions are changed randomly. The observer is

asked to assess the overall quality of both sequences by inserting a mark on a vertical scale. Fig. 6 shows a section of a typical score sheet. The continuous scales are divided into five equal lengths, which correspond to the normal ITU five-point quality continuous scale.

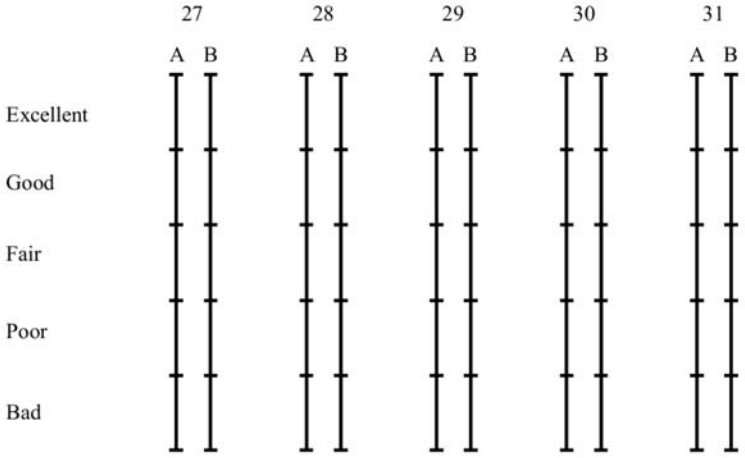


Fig. 6. Continuous quality scale used in DSCQS.

- *Double Stimulus Impairment Scale* (DSIS) – For this method, the reference is always shown before the test sequence and the pair is not repeated. Observers are asked to judge the amount of impairment in the test sequence using a five-level scale. The categories in the scale are 'imperceptible', 'perceptible, but not annoying', 'slightly annoying', 'annoying', and 'very annoying'. This method is adequate for evaluating visible artifacts.
- *Single Stimulus Continuous Quality Evaluation* (SSCQE) – In this method, observers are asked to watch a video (program) of around 20-30 minutes. The content is processed using the conditions under test and the reference is not presented. The observer uses a a slider to continuously rate the quality, as it changes during the presentation. The scale (ruler) goes from 'bad' to 'excellent'.

The most popular assessment procedures of ITU-T Rec. P.910 are:

- *Absolute Category Rating* (ACR) – Also known as Single Stimulus Method (SSM), this method is characterized by the fact that the test sequences are presented one at a time, without the reference. This makes it a very efficient method, compared to DSIS or DSCQS, which have durations of around 2 to 4 times longer. After each presentation, observers are asked to judge the overall quality of the test sequence using a five-level scale. The categories in this scale are 'bad', 'poor', 'fair', 'good', and 'excellent'. A nine-level scale may be used if a higher discriminative power is desired. Also, if additional ratings of each test sequence are needed, repetitions of the same test conditions at different points in time of the test can be used.
- *Degradation Category Rating* (DCR) –This method is identical to the DSIS described earlier.
- *Pair Comparison* (PC) – In this method, all possible pair combinations of all test sequences are shown to viewers, i.e., if there are n test conditions, a total of $n \cdot (n - 1)$

pairs are presented for each reference. The observers have to choose which sequence of the pair he/she thinks has the best quality. This methods allows a very fine distinction between conditions, but also requires a longer period of time when compared to other methods.

Although each assessment method has its own requirements, the following recommendations are valid in most cases:

- The choice of test sequences must take into account the goal of the experiment. The spatial and temporal content of the scenes, for example, are critical parameters. These parameters determine the type and severeness of the impairments present in the test sequences.
- It is important that the set of test scenes spans the full range of quality commonly encountered for the specific conditions under test.
- When a comparison among results from different laboratories is the intention, it is mandatory to use a set of common source sequences to eliminate further sources of variation.
- The test sequences should be presented in a pseudo-random order and, preferably, the experimenter should avoid that sequences generated from the same reference be shown in a subsequent order.
- The viewing conditions, which include the distance from the subject's eye to the monitor and the ambient light, should be set according to the standards.
- The size and the type of monitor or display used in the experiment must be appropriate for the application under test. Callibration of the monitor may be necessary.
- It is best to use the whole screen for displaying the test sequences. In case this is not possible, the sequences must be displayed on a window of the screen, with a 50% grey (Y=U=V=128) background surrounding it.
- Before the experiment starts, the subjects should be tested for visual acuity. After that, written and oral instructions should be given to them, describing the intended application of the system, the type of assessment, the opinion scale, and the presentation methodology.
- At least 15 subjects should be used in the experiment. Preferably, the subjects should not be considered 'experts', i.e., have considerable knowledge in the area of image and video processing.
- Before the actual experiment, indicative results can be obtained by performing a pilot test using only a couple (4-6) of subjects (experts or non-experts).
- A training section with at least five conditions should be included at the beginning of the experimental session. These conditions should be representative of the ones used in the experiment, but should not be taken into account in the statistical analysis of the gathered data. It should be made clear to the observer that the worst quality seen in the training set does not necessarily corresponds to the worst or lowest grade on the scale.
- Include at least two replications (i.e. repetitions of identical conditions) in the experiment. This will help to calculate individual reliability per subject and, if necessary, to discard unreliable results from some subjects.
- Statistical analysis of the gathered data can be performed using standard methods (Snedecor & Cochran, 1989; Hays, 1981; Maxwell & Delaney, 2003; ITU-R, 1998). For each combination of the test variables, the mean value and the standard deviation of the collected assessment grades should be calculated. Subject reliability should also be estimated.

5. Objective video quality metrics

Video quality metrics can be employed to:

- *monitor* video quality;
- *compare* the performance of video processing systems and algorithms; and
- *optimize* the algorithms and parameter settings for a video processing system.

The choice of which type of metric should consider the application and its the requirements and limitations.

In general, video quality metrics can be divided in three different categories according to the availability of the original (reference) video signal:

- *Full Reference* (FR) metric – Original and distorted (or test) videos are available.
- *Reduced Reference* (RR) metric – Besides the distorted video, a description of the original and some parameters are available.
- *No-reference* (NR) metric – Only the distorted video is available.

Figs. 7, 8, and 9 depict the block diagrams corresponding to the full reference, reduced reference, and no-reference video quality metrics, respectively. Observe that on the FR approach the entire reference is available at the measurement point. On the RR approach only part of the reference is available through an auxiliary channel. In this case, the information available at the measurement point generally consists of a set of features extracted from the reference. For the NR approach no information concerning the reference is available at the measuring point.

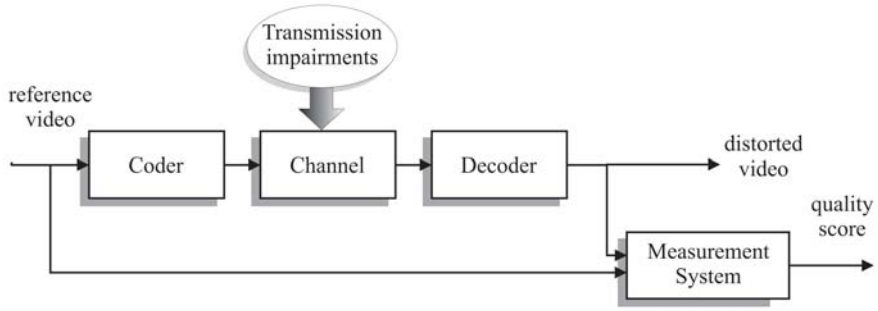


Fig. 7. Block diagram of a full reference video quality assessment system.

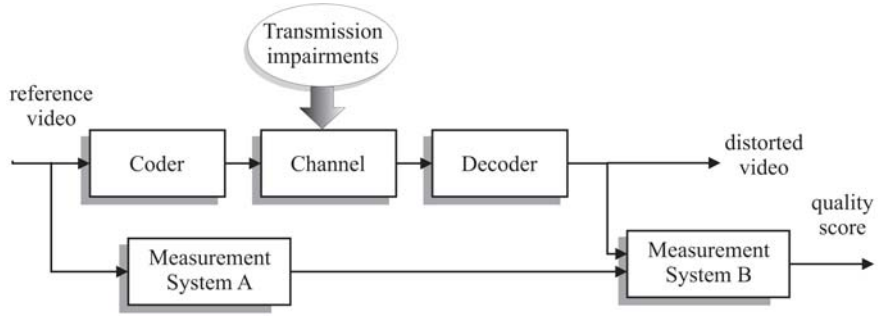


Fig. 8. Block diagram of a reduced reference video quality assessment system.

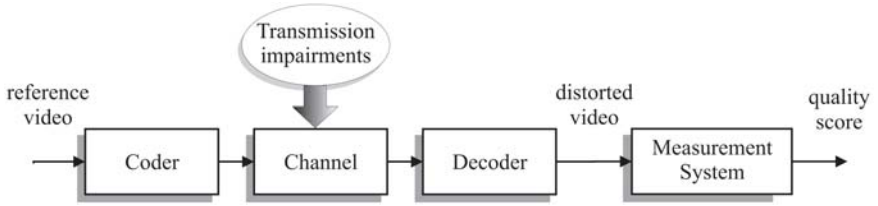


Fig. 9. Block diagram of a no-reference video quality assessment system.

These three classes of metrics are targeted at different applications. FR metrics are more suitable for offline quality measurements, for which a detailed and accurate measurement of the video quality is of higher priority than having immediate results. NR and RR metrics are targeted at real-time applications, where the computational complexity limitations and the lack of access to the reference are the main issues. Comparisons among the performances of several video quality metrics were done by Yubing Wang (Wang, 2006), Eskicioglu and and Fisher (Eskicioglu & Fisher, 1995), Sheikh *et al* (Sheikh et al., 2006), and Avicbas *et al* (Avcibas et al., 2002).

The quality metrics can be classified according to the approach they take for estimating the amount of impairment in a video. There are basically two main approaches. The first one is the *error sensitivity* approach that tries to analyze visible differences between the test and reference videos. This approach is mostly used for full reference metrics, since this is the only type of metric where a pixel-by-pixel difference between the original and test videos can be generated.

The second approach is the *feature extraction* approach that looks for higher-level features that do not belong to the original video to obtain an estimate of the quality of the video. No-reference and reduced reference metrics frequently use the feature extraction approach making use of some a priori knowledge of the features of the original video.

Finally, quality metrics can also be classified according to what type of information they consider when processing the video. Metrics that take into account the how the HVS works are typically called *picture metrics* or perceptual metrics. More simple metrics that only measure the fidelity of the signal without considering its content are called *data metrics*.

In this section, a brief description of a representative set of FR, RR, and NR metrics is presented. Also, a description of data metrics and metrics based on data hiding is presented.

5.1 Data FR fidelity metrics

Data fidelity metrics measure the physical differences between two signals without considering its content. Two of the most popular data fidelity metrics are the mean squared error (MSE) and the peak signal-to-noise ratio (PSNR), which are defined as:

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(X_i - Y_i)^2, \tag{1}$$

and

$$\mathrm{PSNR} = 10 \cdot \log_{10}\frac{255^2}{MSE}, \tag{2}$$

where N is the total number of pixels in the video, 255 is the maximum intensity value of the images, and X_i and Y_i are the i-th pixels in the original and distorted video, respectively.

Strictly speaking, the MSE measures image differences, i.e. how different two images are. PSNR, on the other hand, measures image fidelity, i.e. how close two images are. In both cases, one of the pictures is always the reference (uncorrupted original) and the other is the test or distorted sequence.
The MSE and PSNR are very popular in the image processing community because of their physical significance and of their simplicity, but over the years they have been widely criticized for not correlating well with the perceived quality measurement (Teo & Heeger, 1994; Eskicioglu & Fisher, 1995; Eckert & Bradley, 1998; Girod, 1993; Winkler, 1999). More specifically, it has been shown that simple metrics like PSNR and MSE can only predict subjective rating with a reasonable accuracy, as long as the comparisons are made for the same content, the same technique or the same type of artifact (Eskicioglu & Fisher, 1995).
One of the major reasons why these simple metrics do not perform as desired is because they do not incorporate any HVS features in their computation. In fact, it has been discovered that in the primary visual cortex of mammals, an image is not represented in the pixel domain, but in a rather different manner. The measurements produced by metrics like MSE or PSNR are simply based on a pixel to pixel comparison of the data, without considering what is the content. These simple metrics do not consider, for example, what are the relationships among pixels in an image (or frames). They also do not consider how the spatial and frequency content of the impairments are perceived by human observers.

5.2 Full reference video quality metrics

In general, full reference (FR) metrics have the best performance among the three types of metrics. This is mainly due to the availability of the reference video. Also, since FR are intended for off-line applications, they can be more computational complex and incorporate several aspects of the HVS. The major drawback of the full reference approach is the fact that a large amount of reference information has to be provided at the final comparison point. Also, a very precise spatial and temporal alignment of reference and impaired videos is needed to guarantee the accuracy of the metric.
A large number of FR metrics are *error sensitivity* metrics, which attempt to analyze and quantify the error signal in a way that simulates the human quality judgement. Some examples include the works by Daly (Daly, 1993), Lubin (Lubin, 1995), Teo and Heeger (Teo & Heeger, 1994), Watson (Watson, 1990; 1998;Watson et al., 2001), Van den Branden Lambrecht and Kunt (van den Branden Lambrecht & Kunt, 1998), and Winkler (Winkler, 1999). The group of *full reference* metrics that uses a *feature extraction* approach is much smaller and includes the works of Algazi and Hiwasa (Algazi & Hiwasa, 1993), Pessoa *et al.* (Pessoa et al., 1998), and Wolf and Pinson (Wolf & Pinson, 1999). In this section, we present a brief description of a representative set of full reference video quality metrics.

5.2.1 Visible Differences Predictor (VDP)

The full reference model proposed by Daly (Daly, 1993; 1992) is known as visible differences predictor (VDP). The general approach of the model consists of finding what limits the visual sensitivity and taking this into account when analysing the differences between distorted and reference videos. The main sensitivity limitations (or variations) considered by the model are *light level*, *spatial frequency*, and *signal content*. Each of these sensitivity variations corresponds to one of the stages of the model, as described below:

- Amplitude non-linearity – It is well known that sensitivity and perception of lightness are non-linear functions of luminance. The amplitude non-linearity stage of the VDP

describes the sensitivity variations as a function of the gray scale. It is based on a model of the early retina network.

- Contrast Sensitivity Function (CSF) – The CSF describes the variations in the visual sensitivity as a function of spatial frequency. The CSF stage changes the input as a function of light adaptation, noise, color, accommodation, eccentricity, and image size.
- Multiple detection mechanism – It is modeled with four subcomponents:
  - Spatial cortex transform – It models the frequency selectivity of the visual system and creates the framework for multiple detection mechanisms. This is modeled by a hierarchy of filters modified from Watson's cortex transform (Watson, 1987) that separates the image into spatial levels followed by six orientation levels.
  - Masking function – Models the magnitude of the masking effect.
  - Psychometric function – Describes the details of the threshold.
  - Probability summation – Combines the responses of all detection mechanisms into an unified perceptual response.

A simplified block-diagram of the VDP is depicted in Fig. 10. The output of Daly's metric is a probability-of-detection map, which indicates the areas where the reference and test images differ in a perceptual sense.

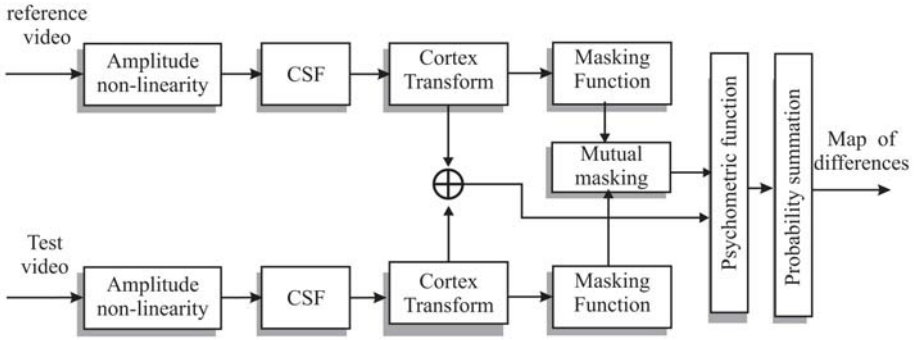


Fig. 10. Block diagram of the visible differences predictor (VDP) (Daly, 1993; 1992).

5.2.2 Sarnoff JND model

The Sarnoff JND model is based on multi-scale spatial vision model proposed by Lubin (Lubin, 1993; 1995). The model takes into account color and temporal variation. Like the metric by Daly, it is designed to predict the probability of detection of artifacts in an image. But, it uses the concept of *just noticeable differences* (JNDs) that are visibility thresholds for changes in images.

The JND unit of measure is defined such that 1 JND corresponds to a 75% chance that an observer viewing the two images detects the difference. JND values above 1 are calculated incrementally. For example, if image A is 1 JND higher than Image B, and image C is 1 JND higher than image A, then image C is 2 JNDs higher than image B. In terms of probability of detection, a 2 JND difference corresponds to 93.75% chance of discrimination, while a 3 JND difference corresponds to 98.44%.

The block diagram of the Sarnoff JND model is depicted in Fig. 11. First, the picture is transformed to the CIE L*u*v* uniform color space (Poynton, 2003). Next, each sequence is filtered and down-sampled using a Gaussian pyramid operation (Burt & Adelson, 1983).

Then, the normalization stage sets the overall gain with a time-dependent average luminance, modelling the HVS insensitivity to overall light level.

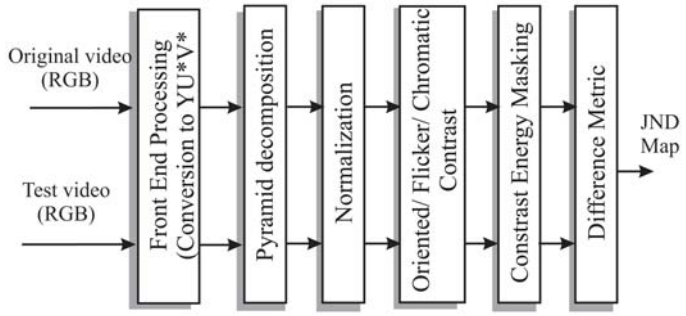


Fig. 11. Block diagram of the Sarnoff JND (Lubin, 1993; 1995).

After normalization, contrast measures are obtained. At each pyramid level, the contrast arrays are calculated by dividing the local difference of the pixel values by the local sum. The result is, then, scaled to be 1 when the image contrast is at the human detection threshold. This gives the definition of 1 JND, which is used on subsequent stages. This scaled contrast arrays are then passed to the contrast energy masking stage in order to desensitize to image "busyness". Then, test and reference are compared to produce the JND map.

5.2.3 Structural Similarity and Image Quality (SSIM)

The Structural SIMilarity and Image Quality (SSIM) (Wang et al., 2004) is based on the idea that natural images are highly "structured". In other words, image signals have strong relationships amongst themselves, which carry information about the structures of the objects in the scene.

To estimate the similarity between a test image and the corresponding reference, the SSIM algorithm measures the luminance $l(x,y)$, contrast $c(x,y)$, and structure $s(x,y)$ of the test image y and the corresponding reference image x, using the following expressions:

$$l(x,y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1}, \tag{3}$$

$$c(x,y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2}, \tag{4}$$

and

$$s(x,y) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3}, \tag{5}$$

where C_1, C_2, and C_3 are small constants given by $C_1 = (K_1 \cdot L)^2$, $C_2 = (K_2 \cdot L)^2$, and $C_3 = C_2/2$. L is the dynamic range of the pixel values (for 8 bits/pixel gray scale images, L = 255), $K_1 \ll 1$, and $K_2 \ll 1$.

The general formula of the SSIM metric is given by

$$SSIM(x,y) = [l(x,y)]^{\alpha} \cdot [c(x,y)]^{\beta} \cdot [s(x,y)]^{\gamma}, \tag{6}$$

where α, β, and γ are parameters that define the relative importance of the luminance, contrast, and structure components. If $\alpha = \beta = \gamma = 1$, the above equation is reduced to

$$\mathrm{SSIM}(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}. \tag{7}$$

The SSIM has a range of values varying between '0' and '1', with '1' being the best value possible. A block diagram of the SSIM algorithm is depicted in Fig. 12. A study of the performance of SSIM has shown that this simple metric presents good results (Sheikh et al., 2006).

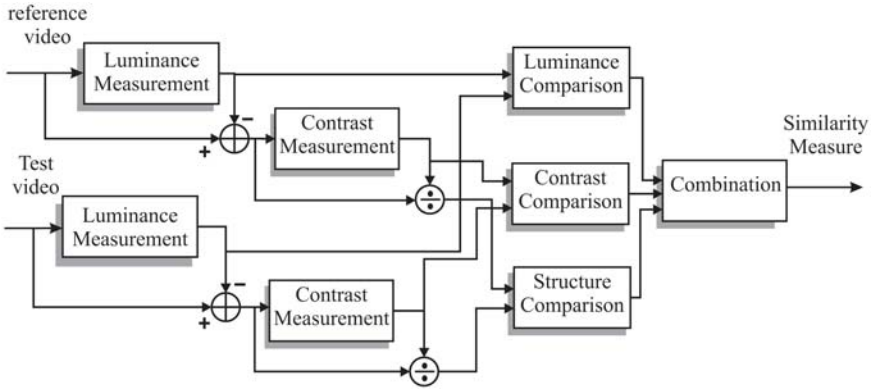


Fig. 12. Block diagram of the SSIM algorithm (Wang et al., 2004).

5.2.4 NTIA Video Quality Metric (VQM)

The video quality metric (VQM) is a metric proposed by Wolf and Pinson from the National Telecommunications and Information Administration (NTIA) (Wolf & Pinson, 1999; Pinson & Wolf, 2004). This metric has recently been adopted by ANSI as a standard for objective video quality. In VQEG Phase II (VQEG, 2003), VQM presented a very good correlation with subjective scores. VQM presented one of the best performances among the competitors.

The algorithm used by VQM includes measurements for the perceptual effects of several video impairments, such as blurring, jerky/unnatural motion, global noise, block distortion, and color distortion. These measurements are combined into a single metric that gives a prediction of the overall quality. The VQM algorithm can be divided into the following stages:

- Calibration – This first stage has the goal of calibrating the video in preparation for the feature extraction stage. With this propose, it estimates and corrects the spatial and temporal shifts, as well as the contrast and brightness offsets, of the processed video sequence with respect to the original video sequence.
- Extraction of quality features – In this stage, the set of quality features that characterizes perceptual changes in the spatial, temporal, and chrominance domains are extracted from spatial-temporal sub-regions of the video sequence. For this, a perceptual filter is

applied to the video to enhance a particular type of property, such as edge information. Features are extracted from spatio-temporal (ST) subregions using a mathematical function and, then, a visibility threshold is applied to these features.

- Estimation of quality parameters – In this stage, a set of quality parameters that describe the perceptual changes is calculated by comparing features extracted from the processed video with those extracted from the reference video.
- Quality estimation – The final step consists of calculating an overall quality metric using a linear combination of parameters calculated in previous stages.

5.3 Reduced reference video quality metrics

Reduced reference (RR) video quality metrics require only partial information about the reference video. To help evaluate the quality of the video, certain features or physical measures are extracted from the reference and transmitted to the receiver as a *side information*. One of the interesting characteristics of RR metrics is the possibility of choosing the amount of side information. In practice, the exact amount of information will be dictated by the characteristics of the side channel that is used to transmit the auxiliary data or, similarly, by the available storage to cache them. Bit rates of the reduced-reference channel can go from zero (for no-reference metrics) to 15 kbps, 80 kbps, or 256 kbps (VQEG, 2009).
Metrics in this class may be less accurate than the *full reference* metrics, but they are also less complex, and make real-time implementations more affordable. Nevertheless, synchronization between the original and impaired data is still necessary. Works in this area include the work by Webster *et al.* (Webster et al., 1993), Brètillon *et al.* (Bretillon et al., 2000), Gunawan and Ghanbari (Gunawan & Ghanbari, 2005), and the work by Carnec *et al.* (Carnec et al., 2003). In this section, we describe the RR metrics by Webster *et al.* and Gunawan and Ghanbari.

5.3.1 Objective video quality assessment system based on human perception

One of the earliest *reduced reference* metrics was proposed by Webster *et al.* (Webster et al., 1993). Their metric is a *feature extraction* metric that estimates the amount of impairment in a video by extracting localized spatial and temporal activity features using especially designed filters. The block-diagram of this metric is depicted in Fig. 13.

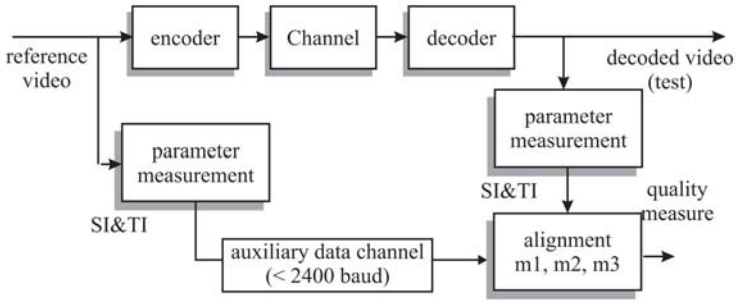


Fig. 13. Block diagram of Webster's algorithm (Webster et al., 1993).

The spatial information (SI) feature corresponds to the the standard deviation of edgeenhanced frames, assuming that degradation will modify the edge statistics in the frames. The temporal information (TI) feature corresponds to the standard deviation of

difference frames, i.e., the amount of perceived motion in the video scene. Three comparison metrics are derived from the SI and TI features of the reference and the distorted videos. The metrics for the reference video are transmitted over the RR channel. The size of the RR data depends upon the size of the window over which SI & TI features are calculated.

5.3.2 Local Harmonic Strength (LHS) metric

The work by Gunawan and Ghanbari (Gunawan & Ghanbari, 2005) proposes a RR video quality metric that is based on a *local harmonic strength* (LHS) feature. The harmonic strength can be interpreted as a spatial activity measure, estimated in terms of vertical/horizontal edges of the picture. In summary, the quality measure is based on harmonic gain and loss estimates obtained from a discriminative analysis of the LHS feature computed on gradient images. A simplified block diagram of the LHS RR video quality metric is depicted in Fig. 14.

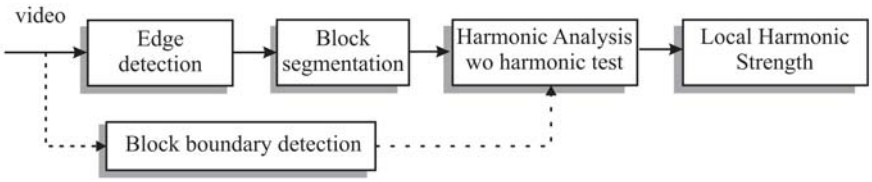


Fig. 14. Block diagram of the LHS algorithm stage of the metric (Gunawan & Ghanbari, 2008; 2005).

The first step of the algorithm is a simple edge-detection stage (3 × 3 Sobel operator) that generates a gradient image. This resulting image is, then, processed by a non-overlap block segmentation algorithm, with a block size large enough to account for any vertical or horizontal activity within the blocks (typically 32 × 32 pixels). An optional alignment with the DCT blocks may also be included to increase the precision of the algorithm.

The LHS is calculated as the accumulated strength of the harmonic frequencies of the blocks on the pictures. Since non-overlapped blocks on a picture are also identified by their spatial location, the collected features from all blocks are organized as a matrix. This matrix has a size which is 32 × 32 smaller than the full resolution picture. LHS matrix features from test and reference pictures are computed separately, and then compared to each other.

The harmonical analysis is performed on the segmented blocks of the gradient image. For this step, a 2-D Fast Fourier Transform (FFT) is applied to each block. The resulting image reveals the appearance of frequency components at certain interval along the two principle axes (horizontal and vertical axes). These frequencies are known as harmonics. The harmonic analysis, then, isolates and accumulates the harmonic components of the resulting FFT spectrum, estimating the value of the LHS feature.

The discriminative analysis is performed after all features from the reference and test images are calculated. Applied to the blocks, the analysis will differentiate between an increase (gain) or decrease (loss) in their strengths, giving an insight on how degradations are distributed over the frame/image. The two quality features produced, namely *harmonic gain* and *harmonic loss*, correspond to blockiness and blurriness on the test image, respectively. To produce a single overall quality measure, spatial collapsing functions (e.g. arithmetic average) are used.

In a more recent algorithm, the author improved the performance of the algorithm by compressing the side information (Gunawan & Ghanbari, 2008), what results in a reduction of the amount of data that needs to be transmitted or stored.

5.4 No-reference video quality metrics

Requiring the reference video or even a small portion of it becomes a serious impediment in many real-time transmission applications. In this case, it becomes essential to develop ways of blindly estimating the quality of a video using a *no-reference* video quality metric. It turns out that, although human observers can usually assess the quality of a video without using the reference, designing a *no-reference* metric is a very difficult task. Considering the difficulties faced by the *full reference* video quality metrics (Eskicioglu & Fisher, 1995; Martens & Meesters, 1997; Rohaly, 2000; VQEG, 2003), this is no surprise.

Except for the metric by Gastaldo *et al.* that uses a neural network (Gastaldo et al., 2001),most of the proposed metrics are *feature extraction* metrics that estimate features of the video. Due to the difficulties encountered in designing NR reference metrics, several metrics rely on one or two features to estimate quality. In most cases, the features used in the algorithms are *artifact signals*, with the most popular being blockiness, blurriness, and ringing. For example, the metrics by Wu *et al.* and Wang *et al.* estimate quality based solely on a blockiness measurement (Wang et al., 2000; Wu & Yuen, 1997; Keimel et al., 2009). The metrics by Farias and Mitra (Farias & Mitra, 2005) and by Caviedes and Jung (Caviedes & Jung, 2001) use four and five artifacts, respectively. In this section, we describe the metrics by Farias and Mitra (Farias & Mitra, 2005) and Oprea *et al.* (Oprea et al., 2009).

5.4.1 No-reference video quality metric based on artifact measurements

As a represention of the metrics based on artifact measurements, we will describe the algorithm proposed by Farias and Mitra (Farias & Mitra, 2005). This approach is based on the assumption that the perceived quality of a video can be affected by a variety of artifacts and that the strengths of these artifacts contribute to the overall annoyance (Ahumada & Null, 1993).

The multidimensional approach requires a good knowledge of the types of artifacts present in digital videos and an extensive study of the most relevant artifacts. The authors performed a series of psychophysical experiments to understand how artifacts depend on the physical properties of the video and how they combine to produce the overall annoyance.

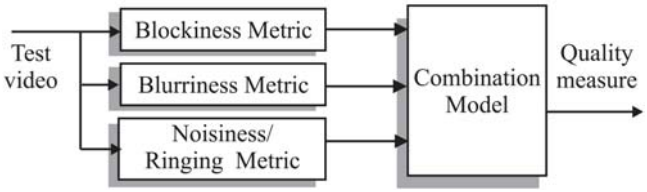


Fig. 15. Block diagram of the *no-reference* metric proposed by Farias and Mitra (Farias & Mitra, 2005).

The block diagram of the metric is as depicted in 15. The algorithm is composed by a set of three artifact metrics (artifact physical strength measurements) for estimating *blockiness*, *blurriness*, *ringing/noisiness*. The metrics are simple enough to be used in real-time applications, as briefly described below.

- The blockiness metric is a modification of the metric by Vlachos (Vlachos, 2000). It estimates the blockiness signal strength by comparing the cross-correlation of pixels inside (intra) and outside (inter) the borders of the coding blocking structure of a frame.

- The blurriness metric is based on the idea that blur makes the edges larger or less sharp (Marziliano et al., 2004; Lu, 2001; Ong et al., 2003). The algorithm measures blurriness by estimating the width of the edges in the frame.
- The noisiness/ringing metric is based on the work by Lee (Lee & Hoppel, 1989) that uses the well known fact that the noise variance of an image can be estimated by the local variance of a flat area. To reduce the content effect a cascade of 1-D filters was used as a pre-processing stage (Olsen, 1993).

To evaluate the performance of each artefact metric, their ability to detect and estimate the artifact signal strength is tested using test sequences containing only the artifact being measured, artifacts other than the artifact being measured, and a combination of all artifacts. The outputs of the individual metrics are also compared to artifact perceptual strengths gathered from psychophysical experiments. A model for overall annoyance is obtained based on a combination of the artifact metrics using a Minkowski metric.

5.4.2 Perceptual video quality assessment based on salient region detection

A recent work by Oprea *et al.* (Oprea et al., 2009) proposes a video quality metric that weighs the distortion measurements on the perceptual importance of the region where it is located.

The first step of this algorithm is to find which are the perceptually important areas of the video frame. For this, the model estimates key features that attract attention: color contrast, object size, orientation, and eccentricity. The measurement of these features will determine which are the important (or salient) areas, producing a saliency map. It is worth pointing out that extracting saliency from video sequences is a complex task because both the spatial extent and dynamic evolution of regions should be considered.

For the detected *salient* areas, a distortion measure is computed using a specialized no-reference metric. The metric considered by the algorithm is a blurriness metric. The blurriness algorithm is based on previous algorithms that estimate the blur by measuring the width of the edges in a frame (Marziliano et al., 2004).

An experiment performed by the authors revealed that the metric has a correlation of about 85% with subjective scores. The algorithm has, nevertheless, limitations concerning the fact that the saliency maps are calculated for the frames individually.

5.5 Metrics using data hiding

An alternative way of implementing a video quality metrics is to use embedding techniques (Sugimoto et al., 1998; Farias, Mitra, Carli & Neri, 2002; Holliman & Young, 2002). The idea consists of embedding in the original video the necessary information to estimate its quality at the time of display. One example of metrics that uses this approach is the work by Farias *et al* (Farias, Mitra, Carli & Neri, 2002). Fig. 16 depicts the block diagram of this video quality assessment system.

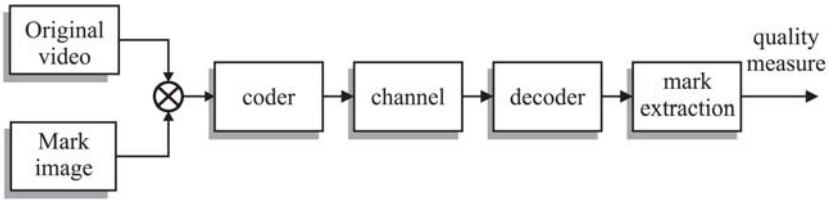


Fig. 16. Block diagram of the video quality assessment system based on data hiding.

At the transmitter, the mark is embedded in each frame of the video using a spread-spectrum technique (Cox et al., 1997). The embedding procedure can be summarized as follows. A pseudo random algorithm is first used to generate pseudo-noise (PN) images $\mathbf{p} = p\,(i, j, k)$, with values -1 or 1 and zero mean. The final mark to be embedded, **w**, is obtained by multiplying the binary image, **m**, by the PN image **p**. Only one binary image is used for all frames, but the PN images vary from frame to frame.
Then, the logarithm of the luminance of the video frame, **y**, is taken and the DCT transform, **LY**, is computed. The mark, **w**, is multiplied by a scaling factor, α, before being added to the luminance DCT coefficients. After the embedding, the DCT coefficients are given by the following expression:

$$LY'(i,j,k) = \begin{cases} LY(i,j,k) + \alpha \cdot w(i,j,k), & 120 \le \mathrm{i} \le 240,\ 120 \le \mathrm{j} \le 240; \\ LY(i,j,k), & \text{elsewhere.} \end{cases}, \tag{8}$$

where i and j are the frequency coordinates and k is the temporal coordinate (index of the frame). For the purpose of assessing the quality of a video, the mark is inserted in the mid-frequencies.
The scaling factor, α, is used to vary the strength of the mark. An increase in the value of α increases the robustness of the mark, but also decreases the quality of the video. After the mark is inserted, the exponential of the video is taken and then the inverse DCT (IDCT). The video is then coded (compressed) and sent over the communication channel.
The process of extracting the mark from the received video is summarized as follows. First, the logarithm of the luminance of the received video, **y**″, is first taken and its DCT calculated. Then, we multiply the mid-frequency DCT coefficients where the mark was inserted by the corresponding pseudo-noise image. Considering that $p(i, j, k) \cdot p(i, j, k) = 1$ because $p(i, j, k)$ is either -1 or +1, we obtain:

$$LY''(i,j,k) \cdot p(i,j,k) = LY(i,j,k) \cdot p(i,j,k) + \alpha \cdot m(i,j), \tag{9}$$

The result of Eq.(9) is then averaged for a chosen number of frames N_f to eliminate the noise (PN signal) introduced by the spread spectrum embedding algorithm. The extracted binary mark is obtained by taking the sign of this average, as given by the following expression:

$$m_r(i,j) = \operatorname{sgn}\left(\frac{1}{N_f} \sum_{k=1}^{N_f} LY(i,j,k) \cdot p(i,j,k) + \alpha \cdot m(i,j,k) \right), \tag{10}$$

Since the PN matrix has zero-mean, the sum $\Sigma_{k=1}^{N_f} LY(i,j,k) \cdot p(i,j,k)$ approaches zero for a large value of N_f. In general, for $N_f \ge 10$ the mark is recovered perfectly, i.e., $\mathbf{m}_r = \mathbf{m}$. When errors are added by compression or transmission, $Y'' = Y' + \eta$ and the extracted mark $\mathbf{m}_r$ is an approximation of **m**. A measure of the degradation of the mark is given by the Total Square Error (TSE) between the extracted mark $\mathbf{m}_r$ and the original binary image:

$$\mathrm{E_{tse}} = \sum_i \sum_j [m(i,j) - m_r(i,j)]^2. \tag{11}$$

The less the amount of errors caused by processing, compression or transmission, the smaller E_{tse} is. On the other hand, the more degraded the video, the higher E_{tse} is. Therefore, the measure given by E_{tse} can be used as an estimate of the degradation of the host video.

6. The work of the video quality experts group

The Video Quality Experts Group (VQEG) was formed in October 1997 in Turin, Italy, to address video quality issues. Since then, it has been conducting formal evaluations of video quality metrics on common test material.

The first task of VQEG was to perform a validation of full reference video quality metrics targeted at TV applications (FR-TV). VQEG outlined, designed and executed a test program to compare subjective video quality evaluations to the predictions of a number of proposed objective metrics (VQEG, 1999). The result of the test was inconclusive (VQEG, 2000). From this first phase, a database of test sequences and their corresponding subjective rating was made available publicly.

In 2003, the second phase of the FR-TV test was completed (VQEG, 2003). The results of these tests have become part of two ITU recommendations (ITU-T, 2004b; a). Contrary to what happened in the first phase, this time the best metrics reached a correlation of around 94% with the subjective scores. The PSNR had a performance of around 70%.

After concluding the FR-TV tests, VQEG conducted a round of tests to evaluate video quality metrics targeted at multimedia applications. The videos considered for this phase had lower bitrates and smaller frames sizes. Besides that, a larger number of codecs and transmission conditions were considered. But, at this first phase the audio signal was not tested. On the 19th of September 2008, the Final Report of VQEGs Multimedia Phase I was released (VQEG, 2008). The correlation results for the submitted FR, RR, and NR metrics were of about 80%, 78%, and 56%, respectively. PSNR had a correlation of around 65%. VQEG has already started working on the second phase of the multimedia tests. In the second phase, both audio and video signals will be tested (simultaneously).

VQEG is currently finishing the tests for evaluation of reduced- and no-reference video quality metrics for television applications ("RR/NR-TV"). The draft report is currently available and the final report should be made available by the time of this publication. VQEG is also working on the evaluation of metrics to be used with High Definition TV (HDTV) content.

One more recent development of VQEG is the test for "hybrid metrics". These metrics estimate quality by looking at not only the decoded video, but also the encoded bitstream. These metrics are targeted at applications like broadcasting, lab applications, live monitoring in network (bitstream) or at end-user (hybrid no reference).

In June 2009, a new activity of VQEG has been launched. It's named Joint Effort Group (JEG) and it consists of an alternative collaborative action. The idea is to work jointly on both mandatory actions to validate metrics (subjective dataset completion and metrics design).

7. Conclusions and new perspectives

In this chapter, we introduced several aspects of video quality. We described the anatomy of the Human Visual System (HVS), discussing a number of phenomena of visual perception that are of particular relevance to video quality. We briefly introduced the main characteristics of modern digital video systems, focusing on the errors (artifacts) commonly present in digital video applications.

We described both objective and subjective methods for assessing video quality. For the subjective methods, we discussed the most common techniques standardized by ITU. We also discussed the main ideas used in the design of objective metric algorithms, listing a

representative set of video quality metrics (FR, RR, and NR). Finally, we discussed the work of the Video Quality Experts Group (VQEG).

It is worth point out that, although much has been done in the last ten years in the area of video quality metrics, there are still a lot of challenges to be solved. Most of the achievements have been in the development of full-reference video quality metrics that evaluate compression artifacts. Much remains to be done, for example, in the area of no-reference and reduced-reference quality metrics.

Also, there has been a growing interest in metrics that estimate the quality of video digitally transmitted over wired or wireless channels/networks. This is due to the popularity of video delivery services over IP networks (e.g. internet streaming and IPTV) or wireless channel (e.g. mobile TV). Another area that has attracted attention is the area of multimedia quality. So far, very few metrics have addressed the issue of simultaneously measuring the quality of all medias involved (e.g. video, audio, text). There is also been an interest in 3D video and HDTV applications.

A new trend in video quality design is the development of *hybrid metrics*, which are metrics that use a combination of packet information, bitstream or decoded video as input (Verscheure et al., 1999; Kanumuri, Cosman, Reibman & Vaishampayan, 2006; Kanumuri, Subramanian, Cosman & Reibman, 2006). The idea here is to also consider parameters extracted from the transport stream and the bitstream (without decoding) in the computation of quality estimation. The main advantage of this approach is the lower bandwidth and processing requirements, when compared to metrics that only consider the fully decoded video.

8. References

Ahumada, A. J., J. & Null, C. (1993). Image quality: a multimensional problem, *Digital Images and Human Vision* pp. 141–148.

Algazi, V. & Hiwasa, N. (1993). Perceptual criteria and design alternatives for low bit rate video coding, *Proc. 27th Asilomar Conf. on Signals, Systems and Computers*, Vol. 2, Pacific Grove, California, USA, pp. 831–835.

Avcibas, I., Avcba, I., Sankur, B. & Sayood, K. (2002). Statistical evaluation of image quality measures, *Journal of Electronic Imaging* 11: 206–223.

Bosi, M. & Goldberg, R. E. (2002). *Introduction to Digital Audio Coding and Standards*, Springer International Series in Engineering and Computer Science.

Bretillon, P.,Montard, N., Baina, J. & Goudezeune, G. (2000). Quality meter and digital television applications, *Proc. SPIE Conference on Visual Communications and Image Processing*, Vol. 4067, Perth,WA, Australia, pp. 780–90.

Burt, P. J. & Adelson, E. H. (1983). The laplacian pyramid as a compact image code, *IEEE Transactions on Communications* COM-31, 4: 532–540.

Carnec, M., Le Callet, P. & Barba, D. (2003). New perceptual quality assessment method with reduced reference for compressed images, *Proc. SPIE Conference on Visual Communications and Image Processing*, Vol. 5150, Lugano, Switzerland, pp. 1582–93.

Caviedes, J. & Jung, J. (2001). No-reference metric for a video quality control loop, *Proc. Int. Conf. on Information Systems, Analysis and Synthesis*, Vol. 13.

Cox, I., Kilian, J., Leighton, F. & Shamoon, T. (1997). Secure spread spectrum watermarking for multimedia, *IEEE Trans. on Image Processing* 6(12).

Daly, S. (1992). The visible difference predictor: An algorithm for the assessment of image fidelity, *Proc. SPIE Conference on Human Vision and Electronic Imaging XII*, p. 2.

Daly, S. (1993). The visible differences predictor: an algorithm for the assessment of image fidelity, *in* A. B.Watson (ed.), *Digital Images and Human Vision*, MIT Press, Cambridge, Massachusetts, pp. 179–206.

de Ridder, H. (1992). Minkowski-metrics as a combination rule for digital-image-coding impairments, *Proc. SPIE Conference on Human Vision, Visual Processing and Digital Display III*, Vol. 1666, San Jose, CA, USA, pp. 16–26.

de Ridder, H. (2001). Cognitive issues in image quality measurement, *Electronic Imaging* 10(1): 47–55.

Eckert,M. & Bradley, A. (1998). Perceptual quality metrics applied to still image compression, *Signal Processing* 70: 177–200.

Eskicioglu, E. & Fisher, P. (1995). Image quality measures and their performance, *IEEE Trans. Image Processing* 43(12): 2959–2965.

Farias, M., Foley, J. &Mitra, S. (2003a). Perceptual contributions of blocky, blurry and noisy artifacts to overall annoyance, *Proc. IEEE International Conference on Multimedia & Expo*, Vol. 1, Baltimore, MD, USA, pp. 529–532.

Farias, M., Foley, J. & Mitra, S. (2003b). Some properties of synthetic blocky and blurry artifacts, *Proc. SPIE Conference on Human Vision and Electronic Imaging*, Vol. 5007, Santa Clara, CA, USA, pp. 128–136.

Farias, M., Foley, J. & Mitra, S. (2004). Detectability and annoyance of synthetic blurring and ringing in video sequences, *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, IEEE, Montreal, Canada.

Farias, M. & Mitra, S. (2005). No-reference video quality metric based on artifact measurements, *Proc. IEEE Intl. Conf. on Image Processing*, pp. III: 141–144.

Farias, M., Mitra, S., Carli, M. & Neri, A. (2002). A comparison between an objective quality measure and the mean annoyance values of watermarked videos, *Proc. IEEE Intl. Conf. on Image Processing*, Vol. 3, Rochester, NY, pp. 469 –472.

Farias, M., Moore, M., Foley, J. & Mitra, S. (2002). Detectability and annoyance of synthetic blocky and blurry artifacts, *Proc. SID International Symposium*, Vol. XXXIII, Number II, Boston, MA, USA, pp. 708–712.

Farias, M., Moore, M., Foley, J. & Mitra, S. (2004). Perceptual contributions of blocky, blurry, and fuzzy impairments to overall annoyance, *Proc. SPIE Conference on Human Vision and Electronic Imaging*, San Jose, CA, USA.

Gastaldo, P., Rovetta, S. & Zunino, R. (2001). Objective assessment of MPEG video quality: a neural-network approach, *Proc. International Joint Conference on Neural Networks*, Vol. 2, pp. 1432–1437.

Girod, B. (1993). What's wrong with mean-squared error?, *in* A. B. Watson (ed.), *Digital Images and Human Vision*, MIT Press, Cambridge, Massachusetts, pp. 207–220.

Grey, H. (1918). Anatomy of the human body. All images are online and are public domain. URL: *http://www.bartley.com/107/*

Gunawan, I. & Ghanbari, M. (2005). Image quality assessment based on harmonics gain/loss information, *Image Processing, 2005. ICIP 2005. IEEE International Conference on*, Vol. 1, pp. I–429–32.

Gunawan, I. & Ghanbari, M. (2008). Efficient reduced-reference video quality meter, *Broadcasting, IEEE Transactions on* 54(3): 669–679.

Haskell, B. G., Puri, A. & Netravali, A. N. (1997). *Digital video: An Introduction to MPEG-2*, Digital multimedia standards, Chapman & Hall: International Thomson Pub., New York, NY, USA.

Hays, W. (1981). *Statistics for the social sciences*, 3 edn, LLH Technology Publishing, Madison Avenue, New York, N.Y.

Holliman, M. & Young, M. (2002). Watermarking for automatic quality monitoring, *Proc. SPIE Conference on Security and Watermarking of Multimedia Contents*, Vol. 4675, San Jose, CA, USA.

ITU (1998). *ISO/IEC 13818: Generic coding of moving pictures and associated audio (MPEG-2).*

ITU (2003). *Recommendation ITU-T H.264: Advanced Video Coding for Generic Audiovisual Services.*

Kanumuri, S., Cosman, P., Reibman, A. & Vaishampayan, V. (2006). Modeling packet-loss visibility in mpeg-2 video, *Multimedia, IEEE Transactions on* 8(2): 341–355.

Kanumuri, S., Subramanian, S., Cosman, P. & Reibman, A. (2006). Predicting h.264 packet loss visibility using a generalized linear model, *Image Processing, 2006 IEEE International Conference on*, pp. 2245–2248.

Keimel, C., Oelbaum, T. & Diepold, K. (2009). No-reference video quality evaluation for highdefinition video, *Acoustics, Speech, and Signal Processing, IEEE International Conference on* pp. 1145–1148.

Klein, S. (1993). Image quality and image compression: A psychophysicist's viewpoint, *in* A. B. Watson (ed.), *Digital Images and Human Vision*, MIT Press, Cambridge, Massachusetts, pp. 73–88.

Lambert, P., de Neve, W., de Neve, P., Moerman, I., Demeester, P. & de Walle, R. V. (Jan. 2006). Rate-distortion performance of h.264/avc compared to state-of-the-art video codecs, *Circuits and Systems for Video Technology, IEEE Transactions on* 16(1): 134–140.

Lee, J. & Hoppel, K. (1989). Noise modeling and estimation of remotely-sensed images, *Proc. International Geoscience and Remote Sensing*, Vol. 2, Vancouver, Canada, pp. 1005–1008.

Lu, J. (2001). Image analysis for video artifact estimation and measurement, *Proc. SPIE Conference on Machine Vision Applications in Industrical Inspection IX*, Vol. 4301, San Jose, CA, USA, pp. 166–174.

Lubin, J. (1993). The use of psychophysical data and models in the analysis of display system performance, *in* A. B. Watson (ed.), *Digital Images and Human Vision*, MIT Press, Cambridge, Massachusetts, pp. 163–178.

Lubin, J. (1995). A visual discrimination model for imaging system design and evaluation, *in* E. Peli (ed.), *Vision models for target detection and recognition*, World Scientific Publishing, Singapore.

Marr, D. (1982). Vision: A Computational Investigation into the Human Representation and Processing of Visual Information, Freeman.

Martens, J. & Meesters, L. (1997). Image dissimilarity, *Signal Processing* 70: 1164–1175.

Marziliano, P., Dufaux, F., Winkler, S. & Ebrahimi, T. (2004). Perceptual blur and ringing metrics: Application to JPEG2000, *Signal Processing: Image Communication* 19(2): 163–172.

Maxwell, S. E. & Delaney, H. D. (2003). *Designing experiments and analyzing data: A model comparison perspective*, Lawrence Erlbaum Associates, Mahwah, NJ.

ITU-R (1998). *Recommendation BT.500-8: Methodology for subjective assessment of the quality of television pictures*.

ITU-T (1999). *Recommendation P.910: Subjective Video Quality Assessment Methods for Multimedia Applications*.

ITU-T (2004a). *Objective perceptual video quality measurement techniques for standard definition digital broadcast television in the presence of a full reference*.

ITU-T (2004b). *Recommendation J.144: Objective perceptual video quality measurement techniques for digital cable television in the presence of a full reference*.

Moore, M. S. (2002). *Psychophysical Measurement and Prediction of Digital Video Quality*, PhD thesis, University of California Santa Barbara.

Olsen, S. I. (1993). Estimation of noise in images: an evaluation, *CVGIP-Graphical Models & Image Processing* 55(4): 319–23.

Ong, E.-P., Lin, W., Lu, Z., Yao, S., Yang, X. & Jinag, L. (2003). No-reference JPEG2000, *Proc. IEEE International Conference on Multimedia and Expo*, Vol. 1, Baltimore, USA, pp. 545– 548.

Oprea, C., Pirnog, I., Paleologu, C. & Udrea, M. (2009). Perceptual video quality assessment based on salient region detection, *Telecommunications, 2009. AICT '09. Fifth Advanced International Conference on*, pp. 232–236.

Pessoa, A. C. F., Falcao, A. X., Silva, A., Nishihara, R. M. & Lotufo, R. A. (1998). Video quality assessment using objective parameters based on image segmentation, *Proc. SBT/IEEE International Telecommunications Symposium*, Vol. 2, Sao Paulo, Brazil, pp. 498–503.

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

Poynton, C. (2003). Digital Video and HDTV - Algorithms and Interfaces, 5th edn, Morgan Kaufmann.

Richardson, I. E. (2003). *H.264 and MPEG-4 Video Compression*, John Wiley & Sons, New York, NY, USA.

Rohaly, A. M. al., e. (2000). Final report from the video quality experts group on the validation of objective models of video quality assessment, *Technical report*, Video Quality Experts Group.

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

Snedecor, G. W. & Cochran, W. G. (1989). *Statistical methods*, 8th edn, Iowa State University Press,

Ames. Sugimoto, O., Kawada, R.,Wada, M. & Matsumoto, S. (1998). Objective measurement scheme for perceived picture quality degradation caused by MPEG encoding without any reference pictures, *Proc. SPIE Conference on Human Vision and Electronic Imaging*, Vol. 4310, San Jose, CA, USA, pp. 932–939.

Teo, P. C. & Heeger, D. J. (1994). Perceptual image distortion, *Proc. IEEE International Conference on Image Processing*, Vol. 2, Austin, TX , USA, pp. 982–986.

van den Branden Lambrecht, C. J. & Kunt, M. (1998). Characterization of human visual sensitivity for video imaging applications, *Signal Processing* 67(3): 255–69.

Verscheure, O., Frossard, P. & Hamdi, M. (1999). User-oriented qos analysis in mpeg-2 video delivery, *Real-Time Imaging* 5(5): 305–314.

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

VQEG (1999). *VQEG subjective test plan (Phase 1)*. ftp://ftp.crc.ca/crc/vqeg/phase1-docs.

VQEG (2000). Final report from the video quality experts group on the validation of objective models of video quality assessment, *Technical report*, http://www.vqeg.org.

VQEG (2003). Final report from the video quality experts group on the validation of objective models of video quality assessment - Phase II, *Technical report*, http://ftp.crc.ca/test/pub/crc/vqeg/.

VQEG (2008). Final report of vqegs multimedia phase i validation test, *Technical report*, http://www.vqeg.org.

VQEG (2009). Rrnr-tv group - test plan draft version 2, *Technical report*, http://www.vqeg.org.

Wang, Y. (2006). Survey of objective video quality measurements, *Technical Report T1A1.5/96- 110*, Worcester Polytechnic Institute.

Wang, Z., Bovik, A. C., Sheikh, H. R., Member, S., Simoncelli, E. P. &Member, S. (2004). Image quality assessment: From error visibility to structural similarity, *IEEE Transactions on Image Processing* 13: 600–612.

Wang, Z., Bovik, A. & Evan, B. (2000). Blind measurement of blocking artifacts in images, *Proc. IEEE International Conference on Image Processing*, Vol. 3, pp. 981–984.

Watson, A. (1987). The cortex transform: rapid computation of simulated neural images, *Computer Vision, Graphics, and Image Processing* 39(3): 311–327.

Watson, A. B. (1990). Perceptual-components architecture for digital video, *Journal of the Optical Society of America, A-Optics & Image Science* 7(10): 1943–54.

Watson, A. B. (1998). Towards a visual quality metric for digital video, *Proc. European Signal Processing Conference*, Vol. 2, Island of Rhodes, Greece.

Watson, A. B., James, H. &McGowan, J. F. (2001). Digital video quality metric based on human vision, *Journal of Electronic Imaging* 10(1): 20–9.

Webster, A. A., Jones, C. T., Pinson, M. H., Voran, S. D. & Wolf, S. (1993). An objective video quality assessment system based on human perception, *Proc. SPIE Conference on Human Vision, Visual Processing, and Digital Display IV*, Vol. 1913, San Jose, CA, USA, pp. 15–26.

Winkler, S. (1999). A perceptual distortion metric for digital color video, *Proc. SPIE Conference on Human Vision and Electronic Imaging*, Vol. 3644, San Jose, CA, USA, pp. 175–184.

Wolf, S. & Pinson,M. H. (1999). Spatial-temporal distortion metric for in-service quality monitoring of any digital video system, *Proc. SPIE Conference on Multimedia Systems and Applications II*, Vol. 3845, Boston, MA, USA, pp. 266–77.

Wolf, S., Pinson, M. H., Voran, S. D. & Webster, A. A. (1991). Objective quality assessment of digitally transmitted video, *Proc. IEEE Pacific Rim Conference on Communications, Computers and Signal Processing*, Victoria, BC, Canada, pp. 477–82 vol.

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

Yuen, M. & Wu, H. R. (1998). A survey of hybrid MC/DPCM/DCT video coding distortions, *Signal Processing* 70(3): 247–78.