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.

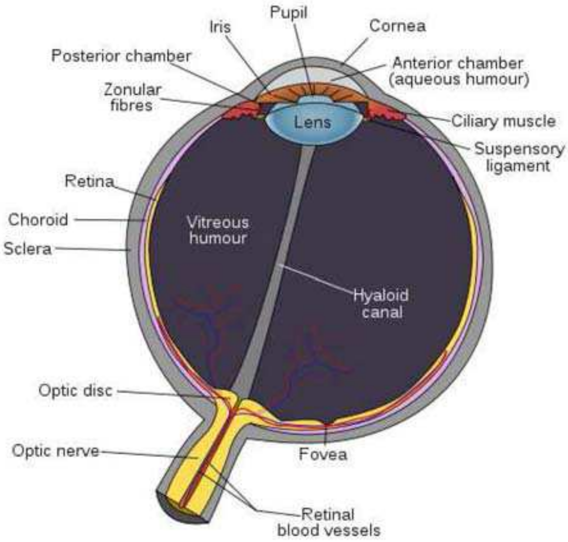


Fig. 1. Transverse section of the human eye (Wikimedia Commons, 2007).

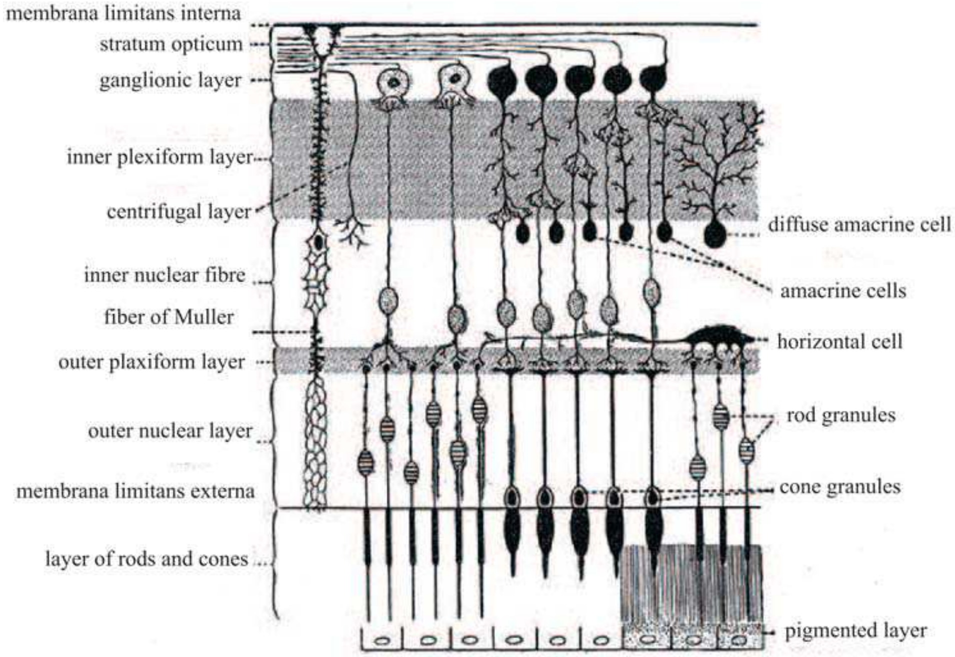


Fig. 2. Plan of retinal neurons. The retina is a stack of several neuronal layers. Light has to pass these layers (from top to bottom) to hit the photoreceptors (layer of rods and cones). The signal propagates through the bipolar and horizontal cells (middle layers) and, then, to the amacrine and ganglion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

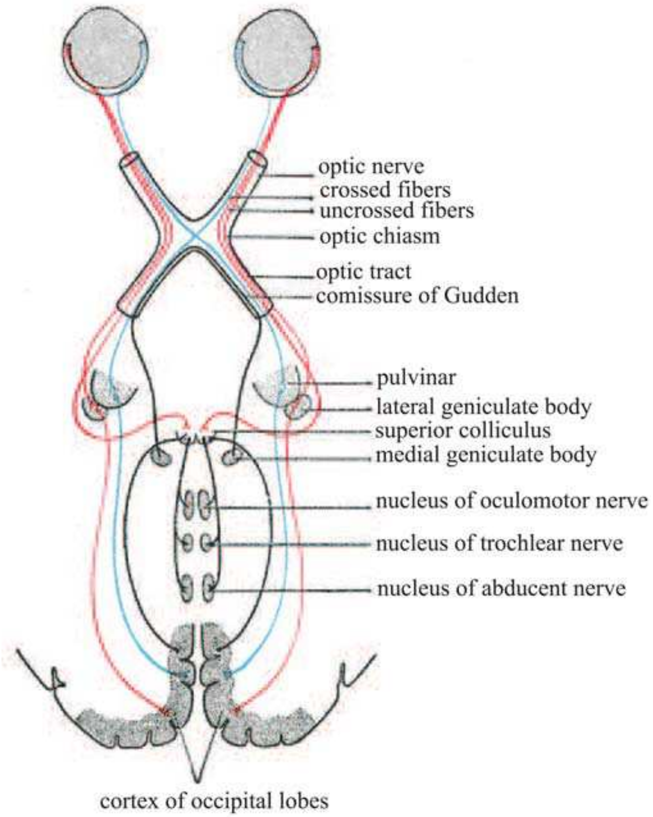


Fig. 3. Scheme showing central connections of the optic nerves and optic tracts. (Adapted from H. Grey (Grey, 1918))

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.

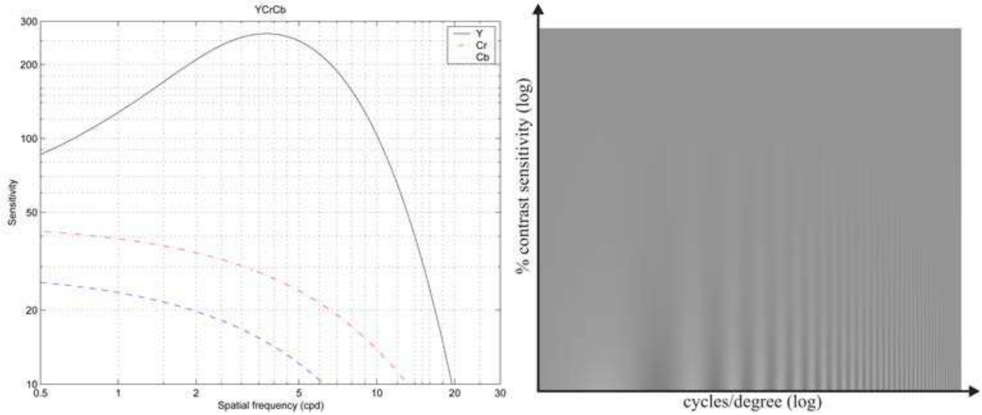
The phenomenon that keeps the contrast sensitivity over a wide range of light intensity is known as Weber's law:

$$\Delta I / I = K$$

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.

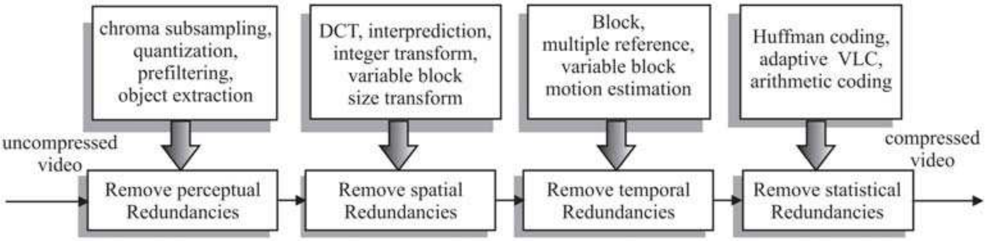


Fig. 5. Functional components in a typical video compression algorithm.

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.

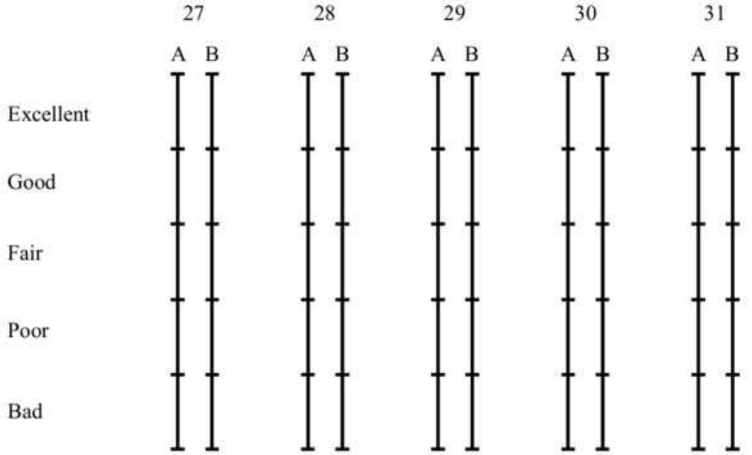


Fig. 6. Continuous quality scale used in DSCQS.

- *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.

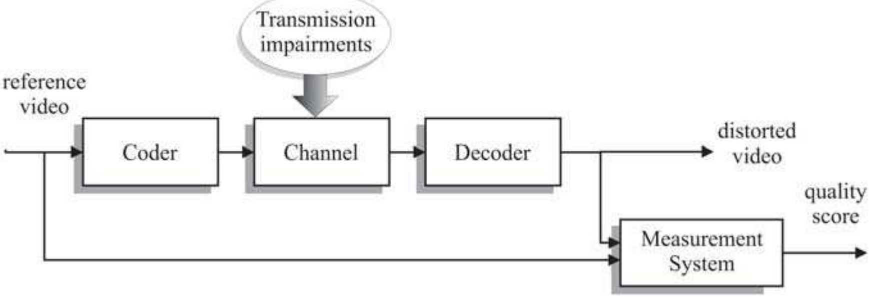


Fig. 7. Block diagram of a full reference video quality assessment system.

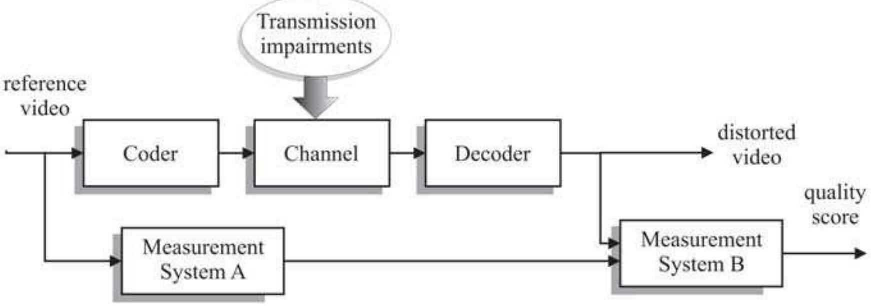


Fig. 8. Block diagram of a reduced reference video quality assessment system.

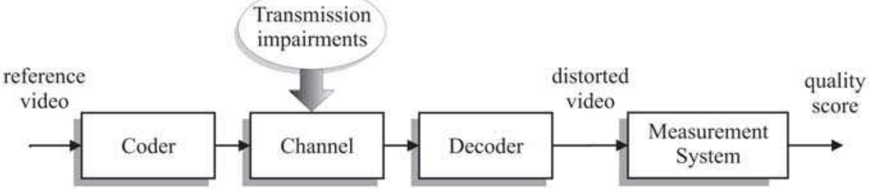


Fig. 9. Block diagram of a no-reference video quality assessment system.

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.

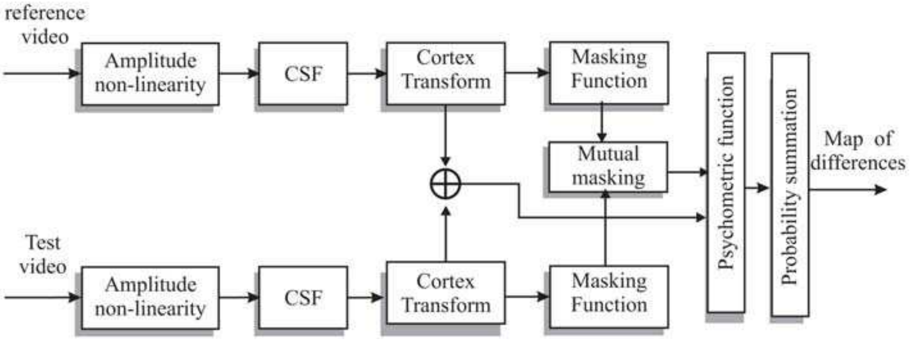


Fig. 10. Block diagram of the visible differences predictor (VDP) (Daly, 1993; 1992).

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).