A Survey of Coded Image and Video Quality Assessment

Amal Punchihewa, Donald G. Bailey and R. M. Hodgson
Institute of Information Sciences & Technology
Massey University, Palmerston North, New Zealand
G.A.Punchihewa@massey.ac.nz, D.G.Bailey@massey.ac.nz, R.M.Hodgson@massey.ac.nz

Abstract

An important aspect of coding is quality assessment. This paper reviews a range of coded image and video quality assessment techniques, their merits, demerits, and their relationships. Objective and subjective assessments are the two major classes of image and video quality assessment. Since picture quality is very subjective, mainly due to the very complex interactions within the Human Visual System (HVS), the primary techniques for measuring picture (both image and video) quality are subjective methods such as ITU Rec. 500-8. Often objective measures of image quality are poorly correlated with perceived picture quality. This makes the development of various objective measures harder. However a few objective quality metrics have been developed that have a high correlation with subjective measures. Researchers in the field still face obstacles when it comes to evaluate the quality of their image and video coding work as no quality metric is universally accepted. The Video Quality Expert Group is working towards such metrics with the participation of many International Telecommunication Union (ITU) study groups.

Keywords: image, video, quality, artefacts, subjective, objective, coding, HVS, VQEG, impairments, survey

1 Introduction

The past years have seen a growing interest in electronic imaging and multimedia applications. Consumer products are appearing on the market that offer a new and broad range of imaging applications to the public. MPEG-2 is in common use. Digital photography, digital TV, photographic quality printers and DVDs are such examples. This new technology has been made possible by recent advances in image capture, image processing and coding. However a difficulty faced by researchers and product developers is that no satisfactory and readily accepted quality metrics available for image and video quality.

For the purpose of communication we often need to acquire, process and deliver visually-presented data, including images and video. A block diagram for this process is shown in Figure 1. Digital image and video systems have many advantages over existing analogue systems. Digital image and video processing and transmission techniques, can be efficient and economical in respect to signal processing, storage and frequency spectrum usage.

When visual information is captured, processed and delivered to the final recipient, the displayed picture may differ from the original. Artefacts are any visible differences that are a direct result of some technical limitation at any stage of the communication process. [1].

In a multimedia environment, an image can be described as a two-dimensional representation of a scene or other visual data. Video can simply be treated as a sequence of images but often contains additional information such as timing or synchronization signals.

The quality of traditional analogue video systems was evaluated using a test pattern or a test signal. For short time periods, the systems are assumed to be as linear and time invariant (LTI). Thus a static test pattern or test signal such as a multi-burst can be used to measure a system attribute such as the frequency response with confidence that the system would respond similarly for other video material (e.g., video with motion). A great deal of research has been performed to relate the traditional analogue video non-linear quality measures (e.g., differential gain, differential phase, short time waveform distortion, etc.) to perceived changes in video quality [2]. These traditional measures of video quality are not applicable in digital video systems where new forms of processing are applied for compression, storage, and transmission. Digital video systems typically adapt and change their behaviour depending upon the input scene properties such as spatial and temporal

Figure 1: Digital Image/Video processing for communication

redundancies to optimise the compression. Most image data contains significant redundancy in the form of high correlation of adjacent pixel values (spatial redundancies), and between adjacent frames of video (temporal redundancies). More recently, digital video systems have been developed that adjust their behaviour based on the available bandwidth. Therefore, to characterise a video or image system, it is important to test for a wide range of potential artefacts for a given system by using an appropriate set of test signals. This paper surveys current status of picture quality assessment of coded digital images and video.

2 Picture Quality and Artefacts

Coding digital images to exploit the redundancy present inevitably results in a loss of quality in the resultant image or video. This is because in removing redundant information, often the some of the essential information is distorted or destroyed, resulting in artefacts, impairments or defects. Frequently the terms impairments, artefacts and defects are used synonymously. However, impairments generally refer to a subjective measure of the degradation of data. Artefacts are visible differences or features that may be added to or removed from an image or video resulting from technical limitation usually of the coding process. Defects may be introduced anywhere within the communication chain, and are therefore more general than artefacts. These influence the human visual system in such a way that the perceived images differ from that intended. Digital video coding introduces fundamentally different types of impairments than those created by traditional analogue techniques.

Artefacts create undesirable visual effects on the picture. There are new forms of artefacts as a result of the migration to the digital domain. The compression algorithms used, the picture content, and the origin of the source material determine the coding artefacts. The greatest technical limitation is the available bandwidth, which affects the compression ratio and data rate. In general, artefacts will become more visible as the compression ratio is increased. Compression-related artefacts such as blockiness and colour bleeding in the digital domain (as depicted in figure 2) are common at low bit rates.

Blocking artefacts are the result of the independent processing of each block in block-based signal processing. Blur and ringing are artefacts result from truncation or quantisation of coefficients in the frequency domain. Ringing artefacts typically appear as sharp oscillations along the edges of an object against a relatively uniform background. Any motion of the object in a video results in these oscillations flickering, giving mosquito noise [1].

Often defects may be decomposed into features or special characteristic. For example, images may report defects with varying strengths of blockiness, loss-of-detail, ringing, and mosquito noise features after MPEG-2 compression [3, 4].

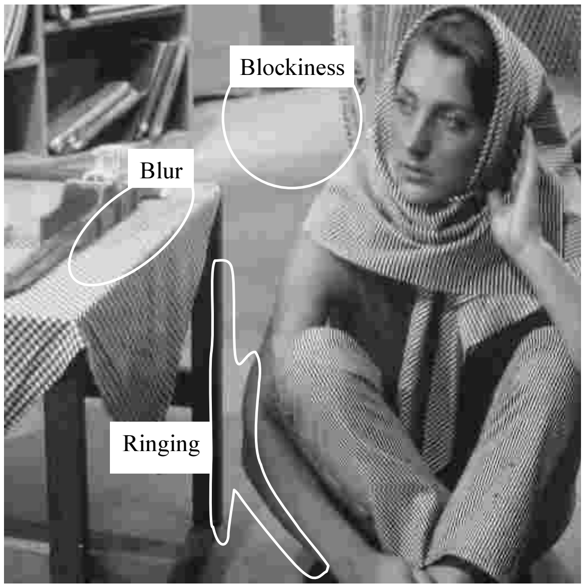

Figure 2: Blockiness, ringing and blur artefacts

For digital video systems, the spatial and temporal information content of the source plays a crucial role in determining the amount of compression that is possible and hence the potential severity of the compression artefacts. The operating characteristics of the digital transmission channel may also be dynamic in relation to bit rate, dropped packets and this will cause the received video quality to vary even when the information content of the input video stream is fixed. When data is transmitted, some of the data may be lost, distorted or reflections may result in multiple data. Noise and distortions can result in missing data blocks or data packets in digital systems (as depicted in figure 3).

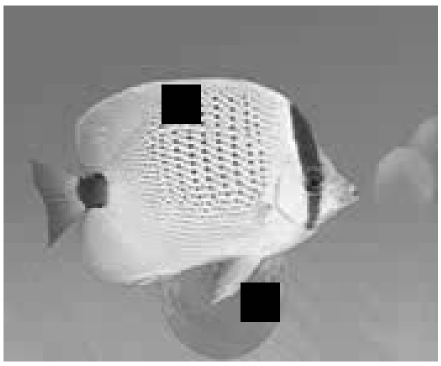

Figure 3: The effects of missing blocks or data packets

Image and video delivery on mobile systems is becoming increasingly common. In low bit-rate video telephony services for mobile third generation systems, channel errors result in highly annoying defects [5]. For these services, it is important to assess the picture as that affects the Quality of Service.

The conventional synchronous model for digital video in which video is reconstructed synchronously at the decoder on a frame-by-frame basis, assumes that transport is delay-jitter-free. However in modern integrated service packet networks such as the Internet, network delay jitter varies over a wide range [6]. The following table 1 summarises the mentioned impairments.

Table 1: Summary of common impairments found in digital image and video systems [1]

Impairment	Description
Blockiness	Distortion of the image characterized by the appearance of an underlying block structure
Blur	Global distortion over the entire image, characterized by reduced sharpness of edges and spatial detail.
Ringing	Appears as echoes of the hard edges in the picture or a rippling of an edge.
Colour bleeding	Occurs at the edges between areas with high colour contrast
Black square blocks	Due to lost packets of data in transmission
Ghosting	Delayed version of the image appear on the screen, to the left or to the right depending on pre or post ghosts

MPEG-4 is a new standard for coding multimedia information, which uses shape coding in place of block coding for objects in images. One purpose of shape coding is to eliminate the artefacts introduced by strong edges with images. No literature could be found about potential new forms of artefacts introduced by shape coding.

3 Classes of Picture Quality Assessment

There are two classes of picture quality assessment methods; subjective and objective.

3.1 Subjective Test Methods

Subjective test methods require human viewers to rate the quality or difference in quality between two image or video clips. In most testing scenarios these two clips differ in that one will be the source and the other will be processed in some manner. Subjective assessment can be a costly and time-consuming process, but one that can yield repeatable results for any given evaluation even when non-expert viewers are used [7]. This type of assessment is particularly necessary in critical situations such as final product (which processes image or video) evaluation and standardization processes where quality must be assured.

Human observers are generally asked to rate video quality in terms of annoyance, where annoyance is a measure of how “bad” the observer thinks the impairment is. The annoyance value correlates with the strength of the impairment (Table 2).

Table 2: Five point scale of ITU Rec. 500-3 subjective assessment of images [8]

Quality	Impairment
5 Excellent	5 Imperceptible
4 Good	4 Perceptible, not annoying
3 Fair	3 Slightly annoying
2 Poor	2 Annoying
1 Bad	1 Very annoying

The Mean Observer’s Score (MOS) is a subjective error measure and is calculated by averaging the annoyance level for all observers. Double stimulus continuous quality scale (DSCQS) is a method in which source and processed video clips are presented in pairs to observers. The video or image presentation sequence is randomised. Viewers grade the quality of each clip then the data is processed in pairs. Until very recent times subjective measures such as MOS and DSCQS offered the most reliable quality measures.

Some issues that arise with subjective assessment include the cost and the fact that these methods cannot be used to monitor video image quality in real time. This process requires special equipment and many people. Traditional analogue objective measurements, while still necessary, are not adequate to measure the quality of systems using digital compression. Thus there is a need to develop new objective methods incorporating the characteristics of the human visual system including perception process [7].

3.2 Objective Test Methods

Objective test methods do not use human subjects, but rather analytically evaluate the video signal. Traditional, analogue methods can accurately measure and assess analogue impairments to a video signal. However, with the introduction and development of digital technologies, visually noticeable artefacts are being manifested in ways that are different from analogue artefacts. This change has led to the need for new objective test methods. The new objective methods analyse the video signal in the image space, utilizing knowledge of the human visual system. These methods implement an algorithm that measures image quality usually based on the comparison of the source and the processed sequences. These algorithms, referred to as models, design to incorporate the functioning of the human visual system while trying to systematically measure the perceptible degradation occurring in the video imagery. In some situations these objective methods may successfully replace the use of subjective assessment.

Due to limitations in the existing quality metrics, it is important to develop quality metrics that are tied to the properties of human visual system. Impairment metrics have been developed based on human visual system models. Lambrecht et al. presented a spatio-temporal vision model and parameterised by psychophysical experiments [9].

Several objective quality metrics have been developed for absolute measurement of picture quality and for use in artefact mitigation algorithms. Attempts were made to quantify the quality without reference to the original source. For example many researchers have proposed algorithms to measure blockiness [6, 10, 11, 12, 13]. Different coding schemes introduce different artefacts making it difficult to design a universal objective quality model. Tan and Ghanbari proposed a multi-metric model comprising of a perceptual model and a blockiness detector. This approach combines a picture quality model for each kind of known distortion according to the perceptual impact of each type of impairment [14].

4 Objective Quality Metrics

The term “quality metric” is generally applied to any physical and psychophysical measure of image quality. Objective measures of impairment are called fidelity measures and are based on the difference between reference and test videos. By definition, fidelity measures require information about the reference video, which can be very restrictive. The fidelity measures are categorized according to the amount of reference information they require. Full-reference fidelity measures require the entire reference video and may not be very practical for real time transmission applications. Reduced-reference fidelity uses a reduced size version of the reference, while non-reference fidelity measures use no reference at all. The difference between the reference and the decoded image provides an objective measure.

4.1 Measurement of Error: Full-reference

Common measures of the errors in signal processing are mean square error (MSE) and the peak signal to noise ratio (PSNR). There are other measures such as total absolute error (TAE), root mean square error (RMS) and signal to noise ratio (SNR).

If the reference signal is x and the y is decoded signal or the approximation of x each having N values, then these metrics are defined as:

$$\mathrm{TAE} = \sum_{i=1}^{N} |(y - x)| \quad (1)$$

$$\mathrm{MSE} = \frac{\sum_{i=1}^{N} (y - x)^2}{N} \quad (2)$$

$$\mathrm{RMS} = \sqrt{\mathrm{MSE}} \quad (3)$$

$$\mathrm{PSNR} = \frac{(\max x)^2}{\mathrm{MSE}} \quad (4)$$

$$\mathrm{SNR} = \frac{\sum x^2}{\sum (x - y)^2} \quad (5)$$

Frequently the SNR or PSNR measures are expressed in decibels. However, none of these measures indicate how the image would appear to the human eye. For example, dithering improves the appearance of certain images but degrades the objective measures.

With digital video, a full reference is often not available at the received end to enable the above quality assessment measures to be used. However, digital techniques enable some data to be embedded with the main stream and this can be used for a partial quality assessment. Two such examples are proposed by Atzori et al. and Carli et al where embedded data within the coded image acts as a reference signal for both error detection and concealment, the process of hiding error [15, 16].

4.2 Double-ended Quality Measure

A dual-ended quality measure compares the input signal with the output signal. To test a compression system, the uncompressed source is compared with the corresponding compressed and then decompressed signal, as shown in Figure 4. This allows the encoding system to be tested in a laboratory environment. However, streams used for the test should be designed or chosen to highlight the details of potential artefacts. Dual-ended measurement also prevents the use of video processing techniques like standards conversion, aspect ratio conversion, digital video effects, fades, between the source and measurement. Finally, the source material requirement makes any dual-ended algorithm impractical for most for day-to-day quality measure operations, especially on multiple channels.

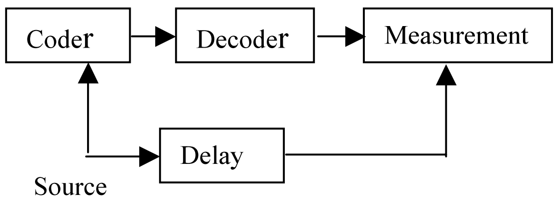

Figure 4: Double-ended picture quality measurements

The compensating delay shown must precisely align the two images to be compared. This can be done by incorporating rugged, easily identifiable regions into the source images or sequences, from which spatial and temporal alignment information can be retrieved even when the sequence has undergone severe degradation. Double-ended measurement is not well suited to monitoring applications where the source is neither available nor controllable, but it is useful as a laboratory tool where those restrictions pose less of a problem [17]. Hence it can be used for testing codec (coder/decoder) design in compression technologies.

4.3 Single-ended Quality Measure

In most practical transmission applications, internet, broadcast, cable, etc., a reference signal is not available unless stored in the system. A single-ended measure only uses the output signal to calculate the quality measure (Figure 5). It is therefore economical and practical, as it doesn't require the input signal. This fact allows measurement to be made of any coded data stream from any point in the communication chain. However it may require certain compression parameters present in the output stream. The researchers who have developed such measures claim that they are accurate measures. Knee [17, 7] has developed a picture appraisal rating (PAR), which is a single-ended picture quality measure for MPEG-2.

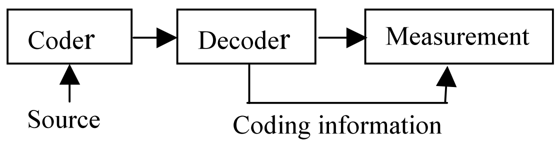

Figure 5: Single ended picture quality measurements

4.4 Commercial Systems

There are only a few products developed by industry to measure image and video quality. Most of these have been optimised for the measurement of MPEG-2 data streams. Rohde and Schwarz have developed a digital video quality analyser and monitor, which functions in real time as a no- reference video quality analysis system [18]. Snell & Wilcox have several products that use the picture appraisal rating (PAR) measure [7]. Tektronix have developed MPEGscope test system, which concentrates more on analysing transmission impairments in MPEG-2 transport streams [19].

5 Video Quality Experts Group (VQEG)

A growing concern of video researchers and broadcasters alike is the assurance and maintenance of an acceptable service quality level for the distribution of video. The Video Quality Experts Group (VQEG) is a group of experts from various backgrounds and affiliations, including participants from several internationally recognized organizations, working in the field of video quality assessment. The group was formed in 1997. With the shift from analogue to digital systems new objective methods are needed to assess the related artefacts. The majority of VQEG participants are active in the International Telecommunication Union (ITU) and VQEG combines the expertise and resources found in several ITU Study Groups to work towards a common goal [20].

The long-standing benchmark used for assessment of video images has been subjective assessment. For more than 20 years, researchers world-wide have used the methods outlined in ITU-R Recommendation 500-8,1998 [8] to evaluate video quality in television services. More recently, the ITU-T has developed Recommendation P.930 to standardize methods for multimedia quality assessment [3]. As noted, because subjective assessment requires the use of human observers to rate video sequences, usually short clips, it is impractical and impossible to use these methods for the in-service continuous evaluation of image quality. Hence, objective methods that emulate the human visual system are required. The first goal of VQEG is to provide the broadcast, telecommunications, and research worlds with objective methods for video image quality evaluation. Several models of objective quality metric are being developed.

6 Discussion and Conclusions

In this review paper, we have presented an overview of image and video quality assessment, techniques and their merits and demerits. Quality assessment is classified into two classes based on the degree of human involvement. Subjective measures are based on HVS and therefore accurately reflect the perceived quality [17]. Objective measures need to be strongly correlated to the HVS and are often targeted to identifying and measuring specific digital coding artefacts.

The simple objective measures of image quality are poorly correlated with perceived picture quality. However more recently developed objective quality metrics, such as picture appraisal rating, have better correlation with subjective measures [17]. All objective quality models developed have their own strengths and weaknesses. The complexities of the human visual system including perception make the goal of designing single objective measure that corresponds to what is gauged by human observers elusive. Other factors which complicate subjective measurements are personal preferences and cultural factors.

As the volume of data and processing power increase, while at the same time the usable spectrum per user is diminishing, it is important that modern digital communication systems handle the available spectrum efficiently. With increasing processing power available at terminal equipment, it is possible to improve the quality of images and video to be displayed by improving the performance of codecs. This is complicated by each new generation of codec introducing different artefacts that weren't apparent in previous generations. However there are no test signals or world standard for an objective measure of picture quality. There is a need for an easy to use portfolio of test signals to be used with a codec to assess its performance.

Researchers in the field still face obstacles when it comes to evaluate the quality of their image and video coding work as no quality metric is universally accepted. The Video Quality Expert Group is working towards such metrics with the participation of many International Telecommunication Union (ITU) study groups.

7 Future work

The authors are developing a set of test signals to be used in the assessment of picture quality of specific codecs. These test signals will be designed to demonstrate the presence of artefacts and to allow that extent to be measured. The test signals are to be developed to be analogous to colour bars and multi-burst signals used in analogue video. These test signals could work with available objective quality metrics or any objective quality metric that the authors may develop in the process. The new test signals are very likely to be codec or standard specific as they will be designed with a knowledge of the specific mechanisms and weaknesses inherent in the codec or standard.

8 References

[1] Punchihewa, A., and Bailey, D.G, Artefacts in Image and Video Systems; Classification and Mitigation, Proceedings of Image and Vision Computing New Zealand 2002, pp197-202 (2002).

[2] Farias, C. Q. de., Thesis dissertation proposal of Mylene C. Q. de Farias, May 2, 2002. “Home page” http://www.engr.ucsb.edu/~mylene, visited on 1/9/2003.

[3] ITU Recommendation P.930, Principles of a Reference Impairment system for Video (1996).

[4] Fenimore, C., Libert, J., and Roitman, P., Mosquito noise in MPEG-compressed video: test patterns and metrics. Proceedings of SPIE: Human Vision and Electronic Image V, San Jose, California, 3959 129-139, (2000).

[5] Dubuc, C., Boudreau D., and Patenaude, F., The Design and Simulated Performance of a Mobile Video Telephony Application for Satellite Third-Generation Wireless System, IEEE Transaction on Multimedia, 3(4) 424-431 (2001).

[6] Chang, Y., Carney, T., Klein, S. A., Messerschmitt, D. G., and Zakho, A., Effects o Temporal Jitter on Video Quality: Assessment Using Psychophysical Methods, Proceedings of the SPIE: Human Vision and Image Processing (1998).

[7] Snell & Wilcox, "Home Page", http://www.snellwilcox.com/, visited on 19/8/2003

[8] ITU Recommendation BT.(500-3 of 1986, 500-8 of 1998). Methodology for the Subjective Assessment of the Quality of Television Pictures.

[9] Lambrecht C. J. Van Den B., AND KUNT, M, Characterization of human visual sensitivity for video imaging applications, Signal Processing, vol. 67, pp. 255-269 (1998).

[10] Bailey, D., Carli, M., Farias, M., and Mitra, S., Quality Assessment for Block-Based Compressed Images and Videos with regard to Blockiness Artefacts, Tyrrhenian International Workshop on Digital Communications Capri, Italy, pp 237-241 (2002).

[11] Farias, M.C.Q., Moore, M. S., Foley, J. M., and Mitra, S. K., Detectability and Annoyance of Synthetic Block and Blurry Video Artefacts, Proceedings of XI European Signal Processing Conference, France, (2000).

[12] Wolf, S., and Pinson, M. H., In-Service Performance Metrics for MPEG-2 Video Systems, Proceedings of made to measure 98 – Measurement Techniques of the Digital Age Technical Seminar, Switzerland, (1998)

[13] Lambrecht, C. J. Van Den B., Verscheure, O., Urbain, J., and Tassin, F., Quality Assessment of Image Features in Video Coding, IEEE International Conference on Image Processing, Santa Barbara, I:302-305 (1995).

[14] Tan, K. T., and Ghanbari, M., A Multi-Metric Objective Picture-Quality Measurement Model for MPEG Video. IEEE Transaction on Circuits and Systems, 10(7) 1208-1213 (2000).

[15] Atzori, L., Natale, F.G.B., and De Perra, C. , A Spatio-Temporal Concealment Technique Using Boundary Matching Algorithm and Mesh-Based Warping (BMA-MBW), IEEE Transaction on Multimedia, 3(3), 326-338 (2001).

[16] Carli, M., Bailey, D., Farias, M., and Mitra, S. K., Error Control and Concealment for Video Transmission using Data hiding, International symposium on wireless personal multimedia communications, Honolulu, Hawaii, pp 812-815 (2002).

[17] Knee, M., The Picture Appraisal Rating (PAR) - a single-ended picture quality measure for MPEG-2, presented at International Broadcast Convention in Amsterdam, (2000).

[18] Rhode and Schowz, "Home Page, http://www.rsd.de/, visited on 19/8/2003

[19] Tektronix, "measurement page", http://www.tektronix.com/Measurement/, visited on 19/8/2003

[20] Video Quality Expert Group, "Home Page" http://www.vqeg.org/, visited on 19/8/2003