CHAPTER

Measuring and Managing Picture Quality

10

CHAPTER OUTLINE

Communicating Pictures. http://dx.doi.org/10.1016/B978-0-12-405906-1.00010-6

Assessing perceptual quality is one of the most critical yet challenging tasks in image and video processing. Visual perception is highly complex, influenced by many confounding factors, not fully understood and difficult to model. For these reasons, the characterization of compression algorithm performance has invariably been based on subjective assessments where a group of viewers are asked their opinions on quality under a range of test conditions. Methods for conducting such trials and for analyzing the results from them are described in Section 10.2. A discussion of the properties of some publicly available subjective test databases is given in Section 10.3.

Objective measures of video quality have conventionally been computed using the absolute or squared difference between the coded version of a picture and its original version. It is however well known that the perceptual distortion experienced by the human viewer cannot be fully characterized using such simple mathematical differences. Because of the limitations of distortion-based measures, perception-based metrics have begun to replace them. These offer the potential for enhanced correlation with subjective opinions, thus enabling more accurate estimates of visual quality. The most promising of these are reviewed in Section 10.4, including Structural Similarity (SSIM), the Video Quality Metric (VQM), Visual Signal-to-Noise Ratio (VSNR), Spatial and Temporal Most Apparent Distortion (MAD), and Motion Tuned Spatio-Temporal Quality Assessment (aka MOVIE). Although these approaches have demonstrated various degrees of improved performance compared to conventional metrics such as Peak Signal-to-Noise Ratio (PSNR), a number of problems including high computational complexity, latency and flexibility of integration into the coding process still need to be resolved. In this context we introduce an alternative, the Perception-based Video Metric (PVM), that successfully addresses some of these issues.

Video metrics are not just used for comparing different coding algorithms. They also play an important role in the picture compression and delivery processes, for example enabling in-loop Rate–Quality Optimization (RQO). Section 10.5 addresses this issue, describing some of the most common techniques that enable us to select the optimum coding parameters for each spatio-temporal region of a video. Finally, methods for controlling the bit rate generated by a codec in the context of time-varying channel conditions (rate control) are described in Section 10.6.

10.1 General considerations and influences

Firstly it is worth stating that assessing the quality of impaired image or video content, whether due to transmission losses or compression artifacts, is not straightforward. It can be achieved in one of two ways:

For colour higher quality versions of Figures 10.14 and 10.15 please refer to the electronic version or the website.

1. **Subjectively:** Requiring many observers and many presentations of a representative range of impairment conditions and content types. Subjective testing conditions must be closely controlled, with appropriate screening of observers and post-processing of the results to ensure consistency and statistical significance. They are costly and time consuming, but generally effective.
2. **Objectively:** Using metrics that attempt to capture the perceptual mechanisms of the human visual system. The main issue here is that simple metrics bear little relevance to the HVS and generally do not correlate well with subjective results, especially at lower bit rates when distortions are higher. More complex, perceptually inspired metrics, although improving significantly in recent years, can still be inconsistent under certain test conditions. The outcome of this is that mean squared error (MSE) based metrics are still the most commonly used assessment methods, both for in-loop optimization and for external performance comparisons.

10.1.1 What do we want to assess?

Good overviews of the reasons and processes for evaluating visual quality are provided in Refs. [1,2]. The primary motivations are summarized below:

1. To compare the performance of different video codecs across a range of bit rates and content types.
2. To compare the performance of different video codecs across a range of channel impairments.
3. To compare the influence of various parameters and coding options for a given type of codec.

The latter is of particular relevance to in-loop Rate–Distortion (or Quality) Optimization RDO (RQO) as we will see later.

10.1.2 Influences on perceived quality

Human visual perception

More than a century ago, vision scientists began to pay attention to our perceptual sensitivity to image and video distortions. We saw in Chapter 2 that this sensitivity varies with screen brightness, local spatial and temporal frequency characteristics, types of motion, eye movements, various types of artifact, and of course the viewing environment. In order to ensure validity of subjective tests and consistency in the performance of objective metrics, the influence of these sensitivities must be, as far as possible, represented in the content used for evaluations and captured in the structure of any metric employed.

It should also be noted that the performance of the HVS varies significantly across subjects, depending on age, illness, fatigue, or visual system defects. It is also possible that biases can influence the opinions of viewers through personal preferences or even boredom.

Viewing environment

In order to provide consistent subjective assessments of picture quality it is essential that the details and parameters of the viewing conditions are recorded and kept as consistent as possible between tests. This is particularly important when comparing results across different laboratories. Several factors about the viewing environment will influence the perception of picture quality. These include:

- **Display size:** The more the display fills peripheral vision, the more engaging it will appear.
- **Display brightness and dynamic range:** Flicker and temporal CSF will depend on display brightness and higher dynamic range displays impart a greater sense of depth.
- **Display resolution:** The spatial and temporal update rates will influence the types of artifacts perceived.
- **Ambient lighting:** High ambient lighting levels and reflections from the display will reduce perceived contrast levels and introduce additional artifacts.
- **Viewing distance:** This parameter interacts with spatio-temporal resolution and will have a significant impact on perceived quality.
- **Audio quality:** Finally it has been clearly demonstrated that variations in audio quality will influence the perception of video quality.

Content type

We have seen in Chapter 2 how content type can influence perceived quality. For example, a plain scene with little motion will be much easier to code than a spatio-temporally busy sequence with complex textures and motions. If codec A is used to code easy content and a similar codec, B, is used to code more difficult content, then it would appear that codec A is superior in terms of rate–distortion performance to codec B, when in reality this is not the case.

Tests should therefore be based on representative and reasonably challenging content offering a range of spatial and temporal activity levels (see Section 10.2.2). To ensure that the focus of the assessor remains on picture quality, rather than on the narrative, short sequences (typically 10 s) are normally selected and these are not usually accompanied by an audio track. All codecs under test should be evaluated on all sequences and at all impairment levels.

Artifact types

We will not list all of the artifact types produced by all types of codec here, but hopefully the reader is by now aware of these from previous chapters. In general they can be classified as follows:

- **Edges:** Due to blocking from transforms or motion estimation, or to contouring effects associated with quantization.
- **Blurring:** Due to the loss of high frequency detail during quantization of block transform or wavelet coefficients.

- **Ringing:** Quantization artifacts in filter bank synthesis stages can cause ringing effects.
- **Dissimilarity:** Due for example to inconsistencies in geometric transformations of planar regions in certain codecs, errors in warping rigid textures or synthesizing dynamic textures (see Chapter 13 for further details).

10.2 Subjective testing

10.2.1 Justification

Despite recent advances in the performance of objective metrics and the large number of them available, none are yet universally accepted as a definitive measure of quality.[1] As a consequence it is necessary to use controlled subjective testing. Subjective assessment methods are employed widely to characterize, compare, or validate the performance of video compression algorithms. Based on a representative set of test content and impairment conditions, they are intended to provide a robust indication of the reactions of those who might view the systems tested.

Most subjective testing experiments conducted are based on recommendations from the ITU and other organizations that have been developed over many years, based on the collective experiences of many organizations. The primary reference documents are ITU-R rec. BT.500 [4] and ITU-T rec. P.910 [5]. The discussion that follows is primarily based on these recommendations.

10.2.2 Test sequences and conditions

Test material

In general, test material should be selected that is appropriate to the problem or application area being addressed. It will normally be defined in terms of its fixed parameters, which include: the number of sequences used, sequence duration, sequence content, spatial resolution, temporal resolution, and the bit depth.

For most general assessments, the test material will be "critical, but not unduly so" [4]. This means that it should contain content that is difficult to code but should remain representative of typical viewing. The sequences selected should always include critical material since results for non-critical material cannot usually be extrapolated. However, if the tests are intended to characterize performance for specific difficult cases, then it would be expected that the material used would be selected to reflect these cases.

In general, it is usual to select at least four types of sequence for testing. This number will provide a minimum coverage of activity levels while not boring the assessors with an over-long test session.

[1] It could be argued that VQM [34], as standardized by ANSI, is more widely accepted than others although it is not the best performing.

Activity or information levels

The amount of spatial and temporal activity present in a clip has a high impact on the degree of compression possible and the quality of the resulting reconstruction. In general, test sequences should be selected that are consistent with the range of channel conditions prevailing in the application areas of interest. It is useful, prior to final selection of test material, to formally assess the spatio-temporal activity levels of the candidate clips to ensure that these cover the appropriate spatio-temporal information space [3,5].

Following the approach recommended in recommendation ITU-T rec. BT.910 [5], the spatial information measure is based on the standard deviation of each frame $\mathbf{S}_z$ in the sequence after Sobel filtering. The maximum value of the standard deviation, over all frames, is selected as the SI metric:

$$\mathrm{SI} = \max_{\forall z} \left\{ \sigma_{\forall(x,y)} \left(\mathrm{Sobel} \left(\mathbf{S}_z(x, y) \right) \right) \right\} \tag{10.1}$$

The temporal information measure (TI) used in BT.910 is based on the standard deviation of the difference, over consecutive frames, between co-located luma pixel values. The TI measure is computed, as above, as the maximum value of the standard deviation over all frames in the sequence. Hence:

$$\mathrm{TI} = \max_{\forall z} \left\{ \sigma_{\forall(x,y)} \left(\mathbf{S}_z(x, y) - \mathbf{S}_{z-1}(x, y) \right) \right\} \tag{10.2}$$

It is usual to produce an SI vs TI plot to ensure that the sequences selected for testing provide adequate coverage of the SI–TI space. An indicative plot, with a good coverage, is shown in Figure 10.1. It should be noted that a slightly different approach to that described above is used by Winkler [3], based on motion vectors. He also includes an information measure based on color.

Test conditions

The *test conditions* refer to those parameters that will change during the test session. These might include codec types, codec parameters, and bit rates, and a typical test would compare two or more codecs or codec configurations for a range of sequences and coding rates.

It is normal to evaluate codec performance over a representative, well-distributed set of test conditions. Most assessment methods are sensitive to the range and distribution of conditions. The sensitivity can be reduced by restricting the range of conditions but also by including some explicit (direct anchoring) extreme cases or by distributing these throughout the test without explicit identification (indirect anchoring). Furthermore, the larger the number of test conditions, the longer the test session becomes. Tests must avoid viewer fatigue and this imposes constraints on the number of sequences employed and the number of test conditions evaluated.

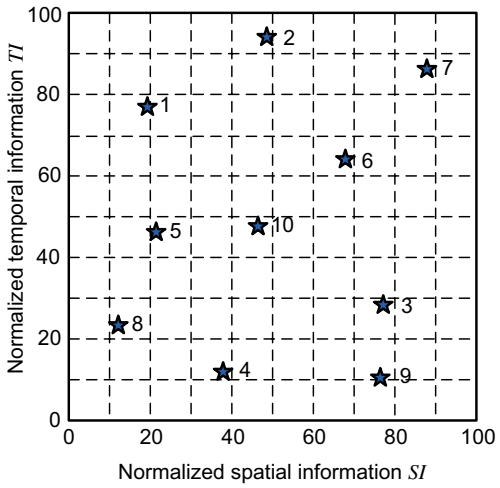


FIGURE 10.1

Spatial and temporal information coverage for a range of test sequences.

The number of presentations is equal to $N = N_s \times N_c \times N_r$ where N_s is the number of source sequences, N_c is the number of test conditions ($N_c = N_{c_1} N_{c_2}$ where N_{c_1} is the number of codecs under test and N_{c_2} is the number of bit rates tested (usually including uncompressed)), and N_r is the redundancy factor (i.e. the number of times each condition is repeated). If each presentation takes T seconds then the time for each test is $N \times T$ seconds and if there are K observers, then the total time for the complete trial is $N \times T \times K$ seconds (not including dead time for change-overs and breaks).

10.2.3 Choosing subjects

Depending on the nature of the test, observers may be expert or non-expert. Studies have found that systematic differences can occur between different laboratories conducting similar tests [4]. The reasons for this are not fully understood, but it is clear that expert observers will view material differently to non-experts. Other explanations that have been suggested include gender, age, and occupation [4]. In most cases, for consumer applications, it is expected that the majority of observers should be non-expert and that none should have been directly involved in the development of the system under test.

Before final selection of the assessors, all candidates should be screened to ensure that they possess normal visual acuity (with or without corrective lenses). This can be done using a Snellen Chart for visual acuity and an Ishihara Chart to check for color vision. The number of assessors used depends on the scale of the test and the sensitivity and reliability of the methodology adopted. Normally it is recommended that at least 15 subjects are employed. This author recommends that the number is slightly higher to allow for outlier removal during results processing.

Table 10.1 Selected subjective testing viewing environment parameters. From Refs. [4, 5].

Viewing condition	Home	Laboratory
Ratio of inactive to peak screen luminance	≤0.02	≤0.02
Display peak luminance	200 cd/m^2	See recommendations BT.814 and BT.815
Background chromaticity	–	D65
Maximum observer angle	30°	30°
Viewing distance	3–6H	3–6H
Room illumination	200 lx	Low (≤ 20 lx)

10.2.4 Testing environment

Testing environments are normally specified according to the requirements of the test—usually meaning either a realistic consumer environment or laboratory conditions:

- **Laboratory conditions:** These are intended to provide test conditions that enable maximum detectability of impairments.
- **Home conditions:** These are intended to enable assessment in viewing conditions closer to typical consumer-end environments.

Selected parameters associated with these two environments are listed in Table 10.1.

10.2.5 Testing methodology and recording of results

A generic diagram of a subjective testing methodology is shown in Figure 10.2. The source delivers the presentation clips, either directly to the subject or via a system under test that introduces impairments dependent on the test conditions set at the time. The subject or subjects will view the content on an assessment display which shows the impaired video clip and in some cases also the unimpaired clip, either simultaneously or sequentially. The controller will control the timing and switching between different test conditions. In some cases the user may be able to influence the order, repetition, and timing of clips.

Testing methodologies broadly fall into two categories—based on their use of single or multiple stimuli. Before we consider these, there are a few general points that are common to all methods.

General principles of subjective testing

- **Duration of test:** To avoid fatigue, BT.500 recommends that the duration of the test session should be limited to 30 min. In practice, it has been the author's experience that tests often do consume more time and 40–45 min is not uncommon, especially in cases where the observer is permitted to control the display and revisit sequences

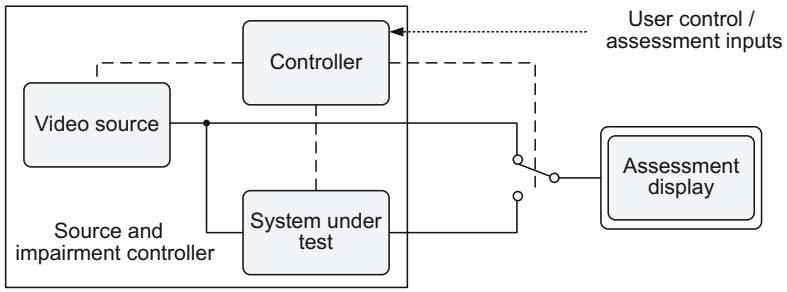


FIGURE 10.2

Generic arrangement for subjective video evaluation.

multiple times (e.g. the Subjective Assessment Methodology for Video Quality (SAMVIQ) [2]).

- **Preparing assessors:** Prior to the test, assessors should receive full instructions on the reason for the test, its content, the types of impairment that will occur and the method of recording results. The first part of any test should include dummy runs which familiarize the assessor(s) with the methodology. These also help to stabilize the observer's opinions.
- **Presentation order:** It is normal during the test session to randomize the presentations—within and across tests. This ensures that any influences of fatigue or adaptation are balanced out. It is also normal to include anchors (extreme cases). In some cases the assessor will know which presentations contain anchors, but this is usually not the case.
- **Recording opinions:** The opinions of each assessor must be recorded for post-test analysis. This is often done using a line bisection process based on gradings such as those shown for the Double Stimulus Continuous Quality Scale test (DSCQS) in Figure 10.3. The observer places a mark on the line for each presentation assessed and this typically falls into one of five quality scale bins as shown in the figure. The final grade is normally scaled to the range 0–100. In other cases, an automated user interface will be provided, such as that shown in Figure 10.4 for the SAMVIQ methodology [2].
- **Recording test conditions:** Because of the inconsistencies that can exist between subjective trials, it is vital to record all salient information about the test and the parameters and equipment used. Good examples are provided by Ebrahimi's group in the context of HEVC—for example, see Refs. [6,7].

Double stimulus methods

Double stimulus evaluations remain the most popular means of evaluating compressed video quality. The most commonly used procedure is ITU-R rec. BT.500 which, although originally intended for television applications, is now used more widely. ITU-T rec. P.910 was targeted specifically at multimedia applications, but shares

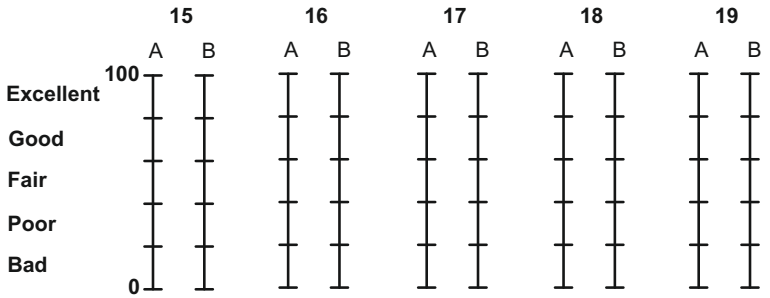


FIGURE 10.3

A section of the DSCQS assessment form.

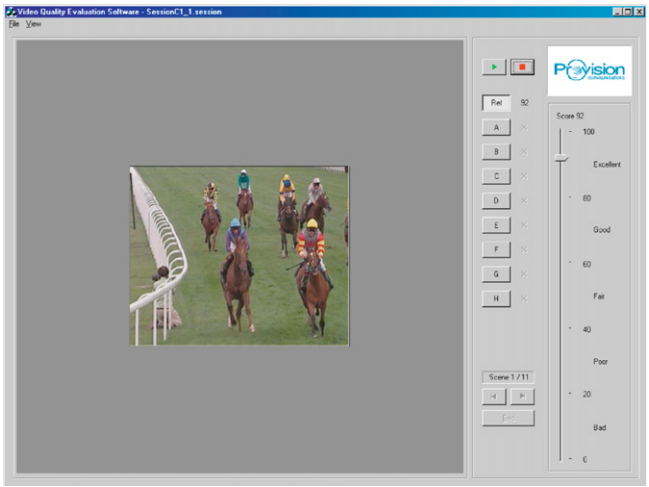


FIGURE 10.4

Interface for SAMVIQ testing. Courtesy: ProVision Communications.

many similarities with BT.500. ITU-R rec. BT.500 describes two main double stimulus methods:

- **Double Stimulus Continuous Quality Scale (DSCQS):** The DSCQS methodology is best suited for cases where the qualities of the test material and the original are similar and the aim is to assess how well the system under test performs relative to the original. The test is arranged as a sequence of paired clips (A and B), one of which is the original (anchor) and the other is the original impaired by the system under test. The assessor does not know which clip is the original as the order is randomized. This is shown in Figure 10.5. The pair is shown twice, the first time for familiarization and the second for voting.

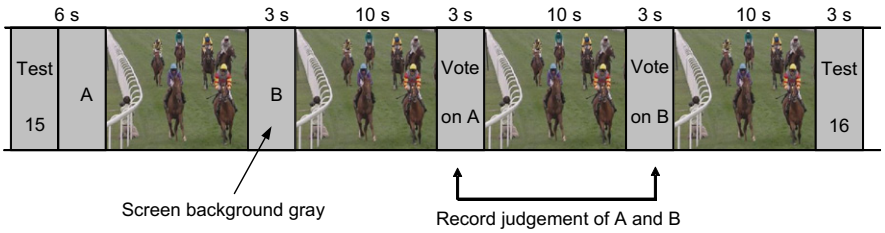


FIGURE 10.5

Typical timing of a DSCQS subjective test.

- **Double Stimulus Impairment Scale (DSIS):** DSIS is similar to DSCQS except that the pair is presented only once and the assessor knows which clip is the anchor as it is always shown before the impaired version. DSIS is generally better suited to assessing the robustness of a system or the effects of more noticeable distortions such as those imparted by transmission errors. This method is very similar to the Degradation Category Rating method (DCR) in ITU-T rec. P.910. The DSIS grading scale again comprises five grades but this time they are labeled:

  – Imperceptible; Perceptible but not annoying; Slightly annoying; Annoying; and Very annoying.

Single stimulus methods

In single stimulus methods, there is no explicit anchor and the assessor is presented with a randomized sequence of test conditions, normally including the original. The Absolute Category Rating (ACR) in ITU-T rec.P.910 does this using 10 s clips across a range of test conditions, with voting on the same scale as in Figure 10.3. A Single Stimulus Continuous Quality Evaluation method is proposed as part of ITU-R rec. BT.500. This is intended to better capture the time and content variations that happen in practical content delivery. Here it is recommended that the test session is organized with longer programme segments (e.g. sport) of around 5 min duration. These are concatenated into a test session of duration 30–60 min comprising a range of program segments with various quality parameters. One of the main advantages of single stimulus methods is reduced testing time.

SAMVIQ [2] is a single stimulus method where the assessor has some control over viewing order and repetitions. The test is organized as a series of test sequences which are assessed in a given order and where the assessor cannot proceed to the next sequence until the previous one has been completely assessed. Within each sequence a number of algorithms and test conditions can be presented in any order as selected by the assessor (however, the selection buttons are randomized for each new sequence evaluated). An example interface with a slider for score entering is shown in Figure 10.4.

Triple stimulus methods

Some authors have proposed that triple stimulus methods provide increased consistency in results, particularly for example when comparing interlaced and progressive scanning methods for HDTV. The Triple Stimulus Continuous Evaluation Scale methodology (TSCES) was proposed by Hoffmann et al. [8], and simultaneously displays the video processed by the system under test alongside (usually they are stacked vertically) two extreme anchors—the original and a low quality coded version. The conventional five-grade scale is also used and the assessor is asked to score the system under test in the context of both anchors.

Pair comparison methods

Pair comparison methods are also double stimulus, but this time are based on comparisons between two systems under test for the same test conditions.

10.2.6 Statistical analysis and significance testing

It is essential that the data collected is analyzed and presented in a robust and consistent fashion. Most test results are scaled to the range 0–100 and that is assumed here. Generally, following the process described in BT.500 [4], let us assume that there are N presentations, where each consists of one of N_c test conditions.

Calculation of mean scores

The mean score across all observers for each presentation is given by:

$$\bar{u}_{csr} = \frac{1}{K}\sum_{k=1}^{K} u_{kcsr} \tag{10.3}$$

where u_{kcsr} is the score for observer k in response to sequence s under test condition c for repetition r, and K is the total number of observers. This is then similarly processed after screening of observers to produce $\bar{u}_{sc}$, the Mean Opinion Score (MOS) for each sequence under a given test condition.

It is common in many tests to compensate scores relative to the reference content. This is particularly important for single stimulus methods where a process of hidden reference removal is applied to produce Difference Mean Opinion Scores (DMOS).

Confidence interval

When the results of a study are presented, it is good practice to also include the confidence interval [9]. The confidence interval for a mean is the interval that will contain the population mean a specified proportion of the time, typically 95%. Confidence intervals are based on the size of each sample and its standard deviation and provide more information than point estimates. The 95% confidence interval is given by:

$$\left[\bar{u}_{csr} - \delta_{csr}, \bar{u}_{csr} + \delta_{csr}\right] \tag{10.4}$$

Algorithm 10.1 Screening of observers in subjective trials using the β_2 test [4].

1. REPEAT for each observer, k:
2. $k = k + 1$;
3. FOR c,s,r = 1,1,1 to C,S,R DO:
4. IF $2 \leq \beta_{2csr} \leq 4$ THEN: IF $u_{kcsr} \geq \bar{u}_{csr} + 2\sigma_{csr}$ THEN $P_k = P_k + 1$; IF $u_{kcsr} \leq \bar{u}_{csr} - 2\sigma_{csr}$ THEN $Q_k = Q_k + 1$;
5. ELSE: IF $u_{kcsr} \geq \bar{u}_{csr} + \sqrt{20}\sigma_{csr}$ THEN $P_k = P_k + 1$; IF $u_{kcsr} \leq \bar{u}_{csr} - \sqrt{20}\sigma_{csr}$ THEN $Q_k = Q_k + 1$;
6. END FOR;
7. IF $\frac{P_k+Q_k}{N_c N_s N_r} > 0.5$ and $\left|\frac{P_k-Q_k}{P_k+Q_k}\right| < 0.3$ THEN reject observer k;
8. UNTIL $k = K$.

where

$$\delta_{csr} = 1.96\frac{\sigma_{csr}}{\sqrt{K}}$$

The standard deviation for each presentation is given by:

$$\sigma_{csr} = \sqrt{\sum_{k=1}^{K} \frac{(\bar{u}_{csr} - u_{kcsr})^2}{(K-1)}} \tag{10.5}$$

The absolute difference between the experimental mean and the true mean (based on an infinite number of observers) is smaller than the 95% confidence interval of equation (10.4) with a probability of 95% (on the condition that the scores are normally distributed).

Screening of observers

The screening (used in DSIS and DSCQS) for a normally distributed set of scores conventionally uses the β_2 test, based on the kurtosis coefficient. When β_2 is between 2 and 4, the distribution can be assumed to be normal. The kurtosis coefficient is defined as the ratio of the fourth order moment to the square of the second order moment. Thus:

$$\beta_{2csr} = \frac{m_4}{(m_2)^2} \quad \text{with} \quad m_x = \frac{\sum_{k=1}^{K}(u_{kcsr} - \bar{u}_{kcsr})^x}{K} \tag{10.6}$$

We now compute P_k and Q_k, as given in Algorithm 10.1, in order to determine whether any observers should be rejected in forming the distribution.

10.2.7 The complete evaluation process

The preceding sections have provided a general overview of subjective testing methods. Both double and single stimulus methods have been formalized through recommendations such as ITU-R rec. BT.500-13, ITU-T rec. P.910 and SAMVIQ [2,4,5].

Algorithm 10.2 Subjective testing methodology.

1. Select the codecs under test and the parameter settings to be used (GOP structures, filtering, RDO modes, ME modes, etc.);
2. Define the coding conditions to be evaluated (spatial and temporal resolutions, bit rates, etc.);
3. Identify or acquire the test sequences that will be used, ensuring that they are sufficiently varied and challenging, representative of the applications of interest and of appropriate duration;
4. Select an appropriate evaluation methodology (dependent on conditions being tested, impairment levels, etc.);
5. Design the test session and environment (complying with recommendations of whichever methodology is being adopted);
6. Dry run the session to remove any bugs or inconsistencies;
7. Organize the full test, inviting sufficient subjects (assessors), with pre-screening to eliminate those with incompatible attributes;
8. Run the tests (including assessor briefing and dummy runs) and collect the results;
9. Perform post-processing to remove outliers and assess significance;
10. Produce test report fully documenting laboratory conditions, test conditions, and analysis/screening methods employed.

The reader is advised to refer to the relevant standard or procedure documents before implementing any actual subjective testing experiments. As a guide, however, a general methodology for subjective testing is described in Algorithm 10.2.

10.3 Test data sets and how to use them

A reliable subjective database has three primary uses. Firstly it will produce a robust comparison of the selected compression methods in the context of the test conditions employed. Secondly it provides a very useful basis for validating objective metrics, based on well-established statistical methods. Finally it can be utilized to characterize HVS properties in the context of compression and hence provides a valuable tool for refining objective quality metrics.

10.3.1 Databases

A brief overview of the primary publicly available databases is presented below. The reader is referred to the appropriate reports associated with each trial for further details. An excellent critique and comparison of currently available databases is provided by Winkler [3].

VQEG FRTV

The earliest major subjective database for objective video quality assessment was generated via the Video Quality Experts Group (VQEG) FRTV Phase 1 program [10]

(followed by Phase 2 in 2003 [11]). The Phase I database was constructed in 2000 to address quality issues associated with the introduction of digital TV standards worldwide. The database used 287 assessors (in four groups) and 20 video sequences with 16 distortion types to generate over 26,000 subjective opinion scores. FRTV-I employed two resolutions, 720 × 486 and 720 × 576, all interlaced. The testing methodology used was the Double Stimulus Continuous Quality Scale (DSCQS) method as described earlier.

Although the FRTV-I database exhibits good coverage and uniformity, it does have limitations in the context of contemporary coding requirements:

1. The artifacts presented do not fully reflect recent advances in video compression as coding is primarily based on MPEG-2.
2. Its assessments are based on standard definition formats. It does not include high definition material.
3. Its content is interlaced, whereas the current trend is toward progressive formats, especially for streaming applications.

However, despite these limitations, because of the large number of sequences, the large number of subjects, its good coverage and its good uniformity [3], VQEG FRTV Phase I is still one of the more commonly used databases for objective quality metric validation.

LIVE

A more recent and equally important subjective test database is that developed in the Laboratory for Image and Video Engineering (LIVE) [12–14]. The LIVE database presents impairments based on both MPEG-2 and H.264/AVC compression algorithms and includes results for simulated wired and wireless transmission errors. Ten reference sequences are used with a resolution of 768 × 432, where each reference has 15 distorted versions. A single stimulus continuous test procedure was adopted for this data set and final scores were again scaled between 0 and 100. Compared to the VQEG FRTV-I database, only 38 assessors were employed in the LIVE viewing trials. According to the analysis of Winkler [3], LIVE also provides a narrower range of source content (in terms of SI and TI) and distortions (in terms of PSNR and MOS values).

Others

Other databases that can be used for assessing metrics include the IRCCyN/IVC [15], the VQEG HDTV database [16], and the VQEG multimedia Phase I database [17]. These are often used for specific research purposes and are in general less suitable for comprehensively validating image and video quality metrics. A further comparison of these can be found in Ref. [3].

10.3.2 The relationship between Mean Opinion Score and an objective metric

Subjective results are frequently used as a benchmark for establishing a relationship between DMOS (or MOS) scores and a specific objective picture quality metric.

The scores produced by the objective video quality metric must be correlated with the viewer scores in a predictable and repeatable fashion. The relationship between predicted and DMOS need not be linear as subjective testing can have non-linear quality rating compression at the extremes of the test range. The linearity of the relationship is thus not so critical, but rather the stability of the relationship and a data sets error-variance that determine predictive usefulness.

BT.500 [4] describes a method of finding a simple continuous relationship between $\bar{u}$ (the mean score) and the metric based on a logistic function. Firstly the range of mean score values is normalized as follows (after screening):

$$p = \frac{(\bar{u} - u_{\min})}{(u_{\max} - u_{\min})} \tag{10.7}$$

Typical relationships between p and a given distortion measure D generally exhibit a skew-symmetric sigmoid form. Hence the function $p = f(D)$ can be approximated by a logistic function of the form:

$$\tilde{p} = \frac{1}{1 + e^{(D - D_M)G}} \tag{10.8}$$

where D_M and G are constants that can be quite simply derived from the experimental data [4].

10.3.3 Evaluating metrics using public (or private) databases

Judgements of the performance of a particular objective metric relative to a body of MOS values associated with a specific database are normally based on certain statistical attributes. These are conventionally related to measures of prediction accuracy, monotonicity, and consistency of its fit. The following measures are commonly used:

Linear correlation

The Pearson Linear Correlation Coefficient (LCC) is used as a measure of the accuracy of fit of the metric to the subjective scores. It characterizes how well the metric under test can predict the subjective quality ratings. The general form is defined for a set of N measurement–prediction pairs (x_i, y_i) as:

$$r_p = \frac{\sum_{i=0}^{N-1}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=0}^{N-1}(x_i - \bar{x})^2}\sqrt{\sum_{i=0}^{N-1}\left(y_i - \bar{y}\right)^2}} \tag{10.9}$$

where x_i would represent the actual MOS (or DMOS) score and y_i would represent the predicted MOS (or DMOS) score.

Rank-order correlation

The degree to which the model's predictions correlate with the relative magnitudes of subjective quality ratings is assessed using a rank-order metric. This characterizes the

prediction monotonicity, i.e. to what degree the sign of differences across tests correlate between the subjective scores and the metric's prediction of them. Conventionally the Spearman Rank-Order Correlation Coefficient is used for this purpose:

$$r_s = \frac{\sum_{i=0}^{N-1} (\mathcal{X}_i - \bar{\mathcal{X}})(\mathcal{Y}_i - \bar{\mathcal{Y}})}{\sqrt{\sum_{i=0}^{N-1} (\mathcal{X}_i - \bar{\mathcal{X}})^2} \sqrt{\sum_{i=0}^{N-1} (\mathcal{Y}_i - \bar{\mathcal{Y}})^2}} \tag{10.10}$$

where $\mathcal{X}_i$ is the rank order of x_i, $\mathcal{Y}_i$ is the rank order of x_i, and $\bar{\mathcal{X}}$ and $\bar{\mathcal{Y}}$ are their median values.

Outlier ratio

The outlier ratio, r_o, effectively measures prediction consistency—how well the metric predicts the subjective scores over the range of content and impairments. An outlier is normally classed as a predicted data point that is greater than a threshold distance from the corresponding MOS point. Conventionally a threshold of twice the standard deviation error of the MOS values is used. So if the number of data points that satisfy equation (10.11) is N_o then the outlier ratio is simply given in equation (10.12).

$$|x_i - y_i| > 2\sigma_{y_i} \tag{10.11}$$

$$r_o = \frac{N_o}{N} \tag{10.12}$$

10.4 Objective quality metrics

10.4.1 Why do we need quality metrics?

Image and video quality assessment plays a crucial role in many aspects of image and video processing, in particular related to video coding and communications. In the context of communicating pictures, they have three primary uses:

1. **Algorithm development and benchmarking:** We have seen previously that subjective evaluations, while very effective in characterizing the strengths and weaknesses of competing coding algorithms, are expensive and time consuming. The existence of reliable and consistent objective metrics provides a much simpler means of comparison. Furthermore they can provide guidance regarding the benefits or otherwise of various algorithmic modifications to a given codec (for example, the benefits of adding a loop filter or using multiple reference frames).
2. **Rate–Quality Optimization:** Quality assessments are increasingly needed in the encoding loop to make instantaneous RQO decisions about which coding modes and parameter settings to use for optimum performance given certain content and rate constraints.

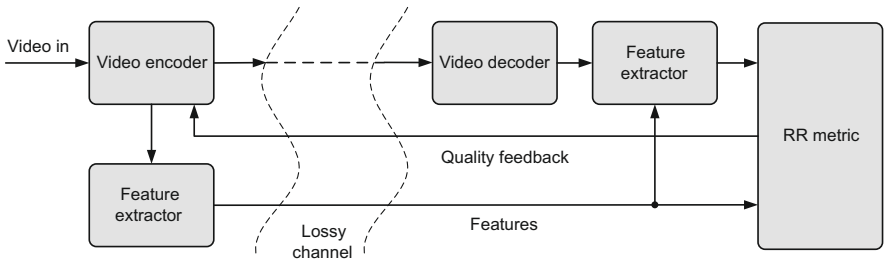


FIGURE 10.6

Reduced reference video quality assessment.

3. **Streaming control:** In the case of network delivery of video content, it is beneficial for the encoder and transmitter to be aware of the quality of the signal at the receiver after decoding. This enables the encoder to be informed of the prevailing channel conditions and hence to make appropriate decisions in terms of rate- and error-control.

Objective quality assessment methods are generally classified as either full-reference (FR), reduced-reference (RR), or no-reference (NR).

- **FR methods** are widely used in applications where the original material is available, such as when assessing image and video coding algorithm performance, or when making Rate–Quality Optimization decisions.
- **NR methods** are normally only employed where reference content is not available [18]. Scenarios could include when evaluating the influence of a lossy communication system at the receiver. It is extremely difficult to produce "blind" metrics and hence the use of NR methods is generally restricted to specific operating scenarios and distortion types. They do not generalize well and reduced reference metrics are preferable if possible.
- **RR methods** [19] use only partial information about the source for predicting quality. They find application in lossy communication networks where quality predictions can be informed by partial knowledge of the original content and possibly also the channel state. Figure 10.6 depicts a possible scenario for an RR metric, where sparse features are extracted at the encoder and transmitted through a protected (low data rate) side channel to the decoder. At the decoder, similar features are extracted from the reconstructed signal and compared in the RR metric. An indication of the reconstruction quality at the decoder can then be fed back to the encoder so that it can make informed coding decisions based on the prevailing channel state. Clearly any additional side information places an overhead on the bit rate of the coded information and this must be assessed in the context of the quality gains achieved.

10.4.2 A characterization of PSNR

For certain types of visual signal under certain test conditions, PSNR can provide a simple and efficient approach to distortion estimation. For example, Huynh-Thu and Ghanbari [20] showed that PSNR can offer consistent results when used to compare between similar codecs or codec enhancements based on the same test data. However, it is equally clear that MSE measures will fail badly for certain impairments that in reality have little perceptual impact. For example, a small spatial or temporal shift, an illumination change or a small variation in a contextual texture [21] will all cause a large increase in MSE. Girod [22] provides a comprehensive overview of the limitations of MSE-based measures and an excellent analysis is also presented by Wang and Bovik [21]. They sum up the situation nicely, listing the assumptions that underpin the use of MSE:

1. Signal quality is independent of temporal or spatial relationships between samples.
2. Signal quality is independent of any relationship between the original signal and the error signal.
3. Signal quality is independent of the signs of the error signal.
4. All samples contribute equally to signal quality.

Based on results from the VQEG FRTV Phase I database, Zhang and Bull [42] analyzed the correlation between PSNR quality predictions (after non-linear regression) and subjective Differential Mean Opinion Scores (DMOS) for a range of different conditions. The primary results are illustrated in Figure 10.7 where it can be observed that in general the points fit only loosely and that the fit worsens at lower quality levels (higher DMOS).

Zhang and Bull investigated how content influences PSNR-based predictions by sub-dividing the FRTV-I dataset into five groups based on the dominant content type (spatial texture, structural elements, plain luminance, dynamic motion or mixed). This grouping was determined subjectively because conventional spatial information (SI) and temporal information (TI) measures do not provide sufficient granularity. The correlations for these content classes are presented in Figure 10.7. It can be observed that, for videos with significant texture content, most subjective DMOS scatter points fall below the PSNR predictions. This illustrates the existence of HVS masking effects for static and dynamic textured content. In contrast, for sequences with significant structural content, the PSNR-predicted DMOS values tend to fall below the subjective scores, indicating that (despite edge masking effects) the HVS is very sensitive to errors around structural elements. The scatter plot for plainer content with little high frequency energy similarly indicates little masking protection.

Zhang and Bull [42] also presented PSNR-predicted DMOS vs DMOS scatter plots based on different types of coding distortion: compressed content, compressed with transmission errors, and compressed with interpolation. These results are shown in the final three sub-plots in Figure 10.7. PSNR can be seen to perform reasonably well for pure coding distortion, where errors are widely distributed, but provides much poorer predictions for the cases with interpolation errors and transmission failures.

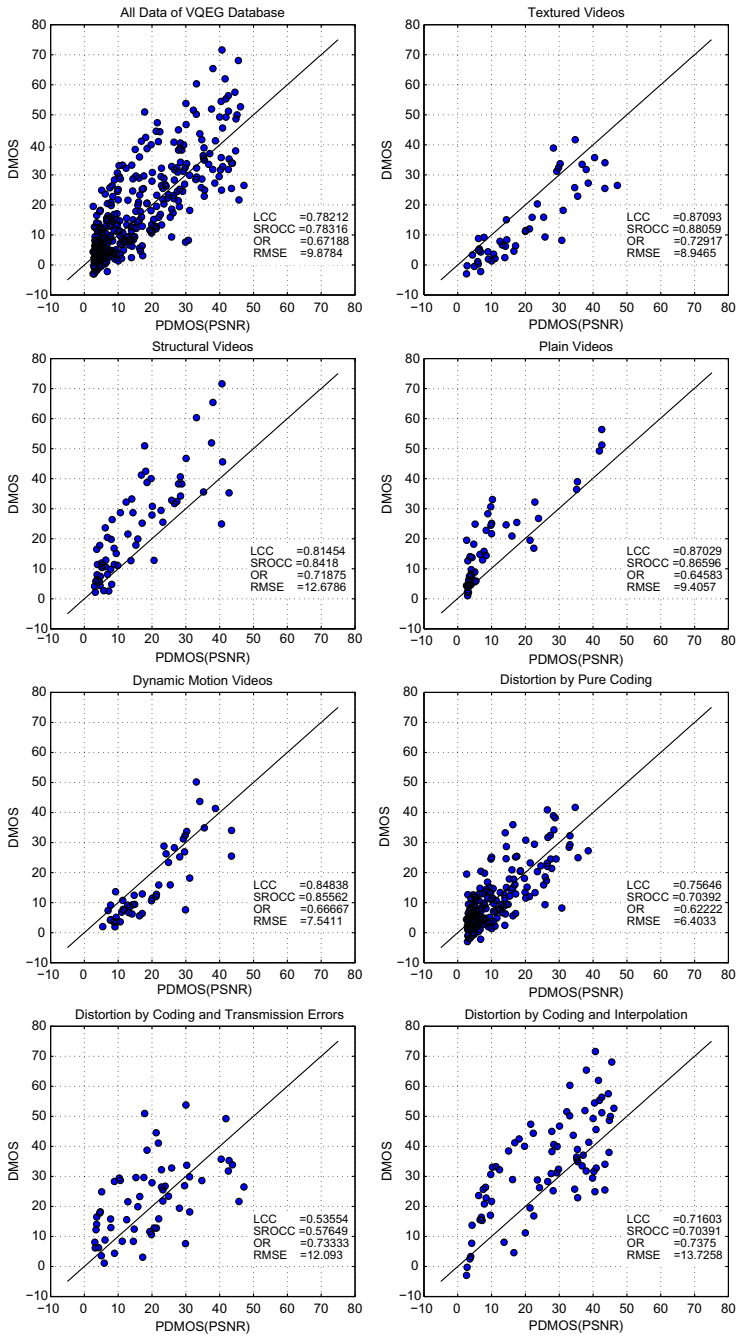


FIGURE 10.7

Scatter plots of DMOS vs PSNR predicted DMOS for various classes of sequence from the VQEG database [42].

It is clear that in the case of highly distorted content, PSNR is not effective and alternative methods are needed.

This analysis confirms that:

1. Visual masking exists and is more evident for content with spatial and temporal textures than for plain luminance areas.
2. Highly visible artifacts such as those in plain areas or due to transmission loss tend to cause the HVS to overestimate distortion.
3. Two distinct perceptual strategies are utilized by the HVS—*near-threshold* and *supra-threshold*. It is clear that artifact detection is more important for cases of lower perceptual quality.

Example 10.1 (Texture masking and PSNR)
Consider the following original 256×256 image (left) and its two distorted versions. In the middle image, the foreground is corrupted by additive Gaussian white noise (AGWN) of variance 0.003. The right-hand image has the background corrupted, again with AGWN of variance 0.003. The PSNR values for the two corrupted images are identical and the foreground and background regions are of equal area.

Compare these images visually and comment on the visibility of the noise (Readers are referred to the electronic version of this figure on the website).

Solution. The PSNR value for the distorted images is 28 dB, so both have identical distortions in the mean squared error sense. Most observers will find the distortions to the background more noticeable than those to the foreground and this is particularly the case where textured content is contextual in the scene and where the content is viewed in the absence of a reference image. This shows that distortions with different perceptual impacts are not always differentiated by PSNR measures.

10.4.3 A perceptual basis for metric development

The principal aim of any video quality metric is to correlate well with visual perception under a wide range of conditions and impairments. Secondary requirements might also include metric complexity and metric latency, particularly relevant for on-line decision making within a codec or transmission system.

HVS characteristics such as contrast sensitivity and visual masking have been exploited, both in coding and in quality assessment. For example, CSF weighting has

been employed to reflect the HVS sensitivity across the range of spatial and temporal frequencies, and visual masking—the reduction of perceived distortions at certain luminances and in spatio-temporal textures—has also been exploited. Masking effects are more evident in spatially textured regions, where the HVS can tolerate greater distortions than in smoother areas. A similar phenomenon exists with dynamic textures.

When HVS properties are exploited, it is generally found that the metric provides enhanced correlation with subjective judgements, compared to conventional distortion measures such as MSE. It has been noted that this improvement is more significant when the distorted content is similar to the original [24].

When distortion becomes significant, visual attention is more likely to be attracted by visible artifacts. Under these circumstances, it is often more efficient to characterize the artifacts rather than to purely compute distortions. Based on this concept, quality metrics have been developed which emulate the perception process using a two-step strategy: *near-threshold* and *supra-threshold*. Examples of this class of metric include those reported by Kayargadde et al. [25], Karunasekera et al. [26], Carnec et al. [27], Chandler and Hemami (VSNR) [28], and Larson and Chandler (MAD) [29].

10.4.4 Perception-based image and video quality metrics

Over the past decade or so, several perceptually inspired objective quality assessment metrics have been proposed. These often exploit visual statistics and features in both frequency and spatial domains. Some of them are described below. For further details the reader is referred to Refs. [1,30–33].

VQM

Pinson and Wolf's VQM [34] is an objective method for video quality assessment that closely predicts subjective quality ratings. It is based on impairment filters that combine measures of blurring, jerkiness, global noise, block distortion and color distortion. VQM computes seven parameters based on different quality features. These are obtained by filtering the impaired and reference videos to extract the property of interest. Spatio-temporal features are then extracted and a quality parameter is obtained for the feature by comparing the statistics of the filtered original and impaired video regions.

The parameters used in VQM are *si_loss* (blurring), *hv_loss* (edge shifts), *hv_gain* (blocking), *si_gain* (edge sharpening), *chroma_spread* (color distribution), *chroma_extreme* (localized color impairments), and *ct_ati_gain* (spatio-temporal impairments). VQM forms a linear combination of these to yield the final quality assessment. In VQEG tests in 2004, VQM demonstrated superior performance in predicting MOS scores compared to all other algorithms evaluated. It has subsequently been adopted as an ANSI and ITU standard.

SSIM

The integrity of structural information in an image or video is an important cue for visual perception. Wang et al. [35] developed an image quality assessment approach,

SSIM (Structural Similarity Image Metric), which estimates the degradation of structural similarity based on the statistical properties of local information between a reference and a distorted image. This is an improved version of the previous Universal Image Quality Index (UIQI) and combines three local similarity measures based on luminance, contrast, and structure. After some simplification, these three terms can be rearranged as in equation (10.13) to give the structural similarity between an impaired image $\mathbf{s}$ and its reference $\mathbf{s}_R$:

$$\text{SSIM} = \frac{(2\mu_s\mu_{s_R} + C_1)\,(2\sigma_{ss_R} + C_2)}{(\mu_s^2 + \mu_{s_R}^2 + C_1)\,(\sigma_s^2 + \sigma_{s_R}^2 + C_2)} \tag{10.13}$$

where μ and σ are the local mean and standard deviation of the two images, C_1 and C_2 are constants used to stabilize the equation in the presence of weak denominators, and σ_{ss_R} is the sample cross correlation (zero mean). Conventionally $C_1 = (k_1 L)^2$ and $C_2 = (k_2 L)^2$ where $k_1 = 0.01$, $k_2 = 0.03$, and L is the dynamic range of the pixel values.

The range of SSIM values extends between -1 and $+1$ and only equals 1 if the two images are identical. It is conventional for SSIM to be calculated using a sliding window, typically of size 11×11, using a circularly symmetric Gaussian weighting function. The SSIM value for the whole image is then computed as the average across all individual windowed results.

The advantage of SSIM is that it offers superior performance to PSNR in many cases and that it is relatively simple to implement. As such it is probably, at the time of writing, the most commonly used non-MSE metric. It has however been recognized that SSIM suffers from a number of problems, particularly that it is sensitive to relative scalings, translations, and rotations. A complex wavelet-based approach, CW-SSIM, has been developed to address these issues [36] as well as an enhanced multiscale version (MS-SSIM) [37]. A further extension to SSIM called V-SSIM which also takes account of temporal information [38] weights the SSIM indices of all frames. This metric has demonstrated improved performance compared to PSNR on the VQEG FRTV Phase I database.

It is interesting to note in the context of Example 10.1 that, although the PSNR values are identical for both distorted images, the SSIM scores are 0.81 for the case of the right-hand image and 0.95 for the middle image. This indicates that SSIM better reflects subjective opinions in the context of texture masking.

MOVIE

Scene motion is a key component in influencing visual perception and it can provide a considerable degree of artifact masking. However, perceived quality is also dependent on viewing behavior; for example, whether or not a moving object is being tracked by the observer. Seshadrinathan et al. [13] introduced a motion tuned spatio-temporal quality assessment method (MOVIE). MOVIE analyzes both distorted and reference content using a spatio-temporal Gabor filter family, and the quality index consists of a spatial quality component (inspired by SSIM) and a temporal quality component based on motion information.

The accuracy of MOVIE-based predictions has been evaluated using the VQEG FRTV Phase I database, where it was demonstrated to offer significant correlation improvements compared to both PSNR and SSIM. It also demonstrates excellent performance on the LIVE video database. One issue with MOVIE however is its high computational complexity, due to the large number of Gabor filters used and the temporal extent required for its calculation.

VSNR

The near-threshold and supra-threshold properties of the HVS were exploited by Chandler and Hemami in their Visual Signal-to-Noise Ratio (VSNR) still image metric [28]. This emulates the cortical decomposition of the HVS using a wavelet transform. A two-stage approach is then applied to assess the detectability of distortions and determine a final measure of visual SNR. VSNR has been evaluated using the LIVE image database with very good results.

MAD and STMAD

Building on the approach used in VSNR, Larson and Chandler [29] developed the Most Apparent Distortion model (MAD). MAD models both near-threshold distortions and appearance-based distortions. It employs different approaches for high quality images (near threshold distortions) and low quality images (supra-threshold distortion). These are combined using a non-linear model to obtain a final quality index.

MAD was extended to cope with temporal components in Ref. [39] where spatio-temporal slices are processed based on the spatial MAD and the results are then weighted using motion information. The temporal MAD index is combined with spatial MAD (computed from individual frames) to obtain spatio-temporal MAD (ST-MAD) for video. Excellent correlation performance with subjective results is reported based on the LIVE video database.

VDP and VDP-2

In VDP-2 Mantiuk et al. [40] built on the previous VDP metric to create a metric that correlates well with subjective opinions for content with extended or diminished intensity ranges. The VDP-2 metric predicts both error visibility and quality (MOS) and was based on new contrast sensitivity measurements. The model was validated using LIVE (image) and TID2008 image databases and demonstrated improved performance compared to its predecessor HDR-VDP and VDP metrics, especially for low luminance conditions. It also compared favorably with the MS-SSIM metric.

Reduced complexity metrics and in-loop assessment

Current image and video quality assessment methods have been developed based on various HVS characteristics in order to achieve quality prediction close to subjective opinions. These approaches provide enhanced correlation with subjective scores

but are often disadvantaged by high complexity or high latency, and hence are not appropriate for real-time operation.

Recently, Soundararajan and Bovik presented a reduced reference video quality metric [19] based on spatio-temporal entropic differencing. This performs well and supports low complexity operation through the use of reduced spatio-temporal content.

In the context of perceptual video coding, Zhang and Bull proposed an Artifact-based Video Metric (AVM) [23] using the DT-CWT as the basis for assessment of both conventionally compressed and synthesized content. AVM correlates well with subjective VQEG scores and has the advantage that it can be easily integrated into a synthesis-based framework because of its high flexibility and low complexity due to extensive parameter reuse.

PVM

A Perception-based Video quality Metric (PVM) was recently proposed by Zhang and Bull [41,42], building on their AVM model. PVM simulates the HVS perception processes by adaptively combining noticeable distortion and blurring artifacts (both exploiting the shift-invariance and orientation selectivity properties of the Dual-Tree Complex Wavelet Transform) using an enhanced non-linear model. Noticeable distortion is defined by thresholding absolute differences using spatial and temporal masks which characterize texture masking effects, and this makes a significant contribution to quality assessment when the distorted video is similar to the original. Blurring artifacts, estimated by computing high frequency energy variations and weighted with motion speed, are found to improve the overall metric performance in low quality cases when it is combined with noticeable distortion.

Importantly PVM, as with its predecessor, AVM, is intended to be used with synthesized as well as conventionally coded material. Early results indicate that these are the only metrics capable of robust performance in this respect.

Comparing results

The performances of PVM, MOVIE, and ST-MAD are compared for both LIVE video and VQEG FRTV Phase I databases as scatter plots in Figure 10.8. Values for correlation coefficients, Outlier ratio and RMSE are given on each graph. It can be seen that PVM is more consistent across the two databases and offers superior performance in the context of the VQEG database.

10.4.5 The future of metrics

While it is clear that perceptual metrics have advanced significantly in the past decade or so, they are still not widely accepted for general picture quality assessment. There are many complex reasons for this. For example, they mostly measure video fidelity (the closeness of the impaired version to the original), they generally measure degradation (i.e. the impaired version is assumed to be worse than the original), and they

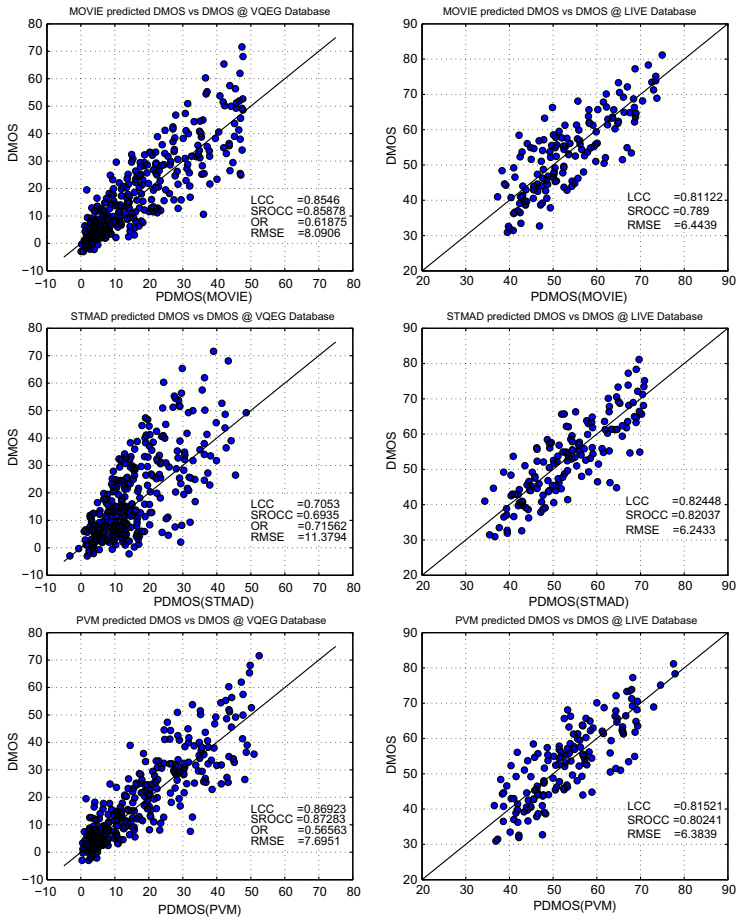


FIGURE 10.8

Comparison of MOVIE, ST-MAD, and PVM on LIVE and VQEG FRTV database [42].

do not generally take account of viewing patterns (viewers tend to focus on regions of interest).

One of the most challenging problems for the future is to create a reliable, yet low complexity, in-loop quality assessment measure with which to precisely estimate subjective quality and detect any noticeable coding artifacts. MSE methods are used extensively for this at present, but will be inappropriate as synthesis-based coding becomes more commonplace.

For real-time or in-loop processing, quality metrics should be able to:

1. Perform assessment at different spatial levels (such as GOP, picture, region, or coding unit).
2. Offer manageable computational complexity.

3. Differentiate the effects of an extended video parameter space, including higher dynamic range, frame rate and resolution, and take account of different environmental conditions.
4. Ideally provide compatibility with emerging perceptual coding and analysis–synthesis compression methods (see Chapter 13).

Few if any existing metrics meet these criteria.

10.5 Rate–distortion optimization

Building on our introduction to rate and distortion in Chapter 4, we will consider this important area of compression in more detail here. As outlined previously, the bit rate for any compressed sequence will depend on a number of factors. These include:

- The video content (high spatio-temporal activity will in general require more bits to code).
- The encoding algorithm used (e.g. wavelet or DCT based, intra-only or motion compensated).
- The encoding parameters selected. At the coarsest level this includes things such as spatial resolution and frame rate. At a finer granularity, issues such as quantizer control, intra vs inter modes, and block size choices will be key. The difference in performance between an encoding based on good parameter choices and one based on poor choices can be very significant.

The rate vs distortion (or rate vs quality) characteristic for a given codec–sequence combination provides a convenient means of comparing codecs and assessing how parameter selections can influence performance. A plot of operating points might, for a range of different parameter values, look like Figure 10.9. The aim is to ensure that the parameters selected at any time are those that yield the lowest possible distortion for any given coding rate. To achieve this we must perform Rate–Distortion (RDO) or Rate–Quality Optimization (RQO).

For further reading on this topic, excellent reviews of rate–distortion theory and practice are provided by Ortega and Ramchandran [43] and Sullivan and Wiegand [44]. Wang et al. [45] also provide a good overview of R-D bounds. For more recent work on RDO related to current standards, the reader is referred to Refs. [46,47].

10.5.1 Classical rate–distortion theory

Rate–distortion theory lies at the heart of source coding. It refers to the trade-off between source fidelity and coding rate and can be stated simply as either:

1. representing a source with the minimum number of bits given a quality constraint, or,
2. achieving the optimum quality for a given bit rate.

Shannon's *separation principle* states that, for the case of transmission of data over a noisy channel (under certain assumptions), the optimal use of bandwidth results if

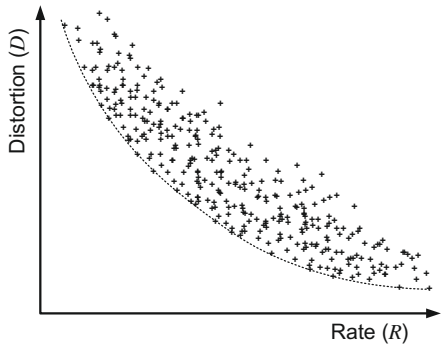


FIGURE 10.9

Rate–distortion plot for various coding parameter choices. The associated convex hull provides the operational R-D characteristic for the given codec–source combination.

source and channel coding are treated independently. In other words, we should first obtain the most efficient data representation for the source in an error-free environment and then add an appropriate level of protection to ensure error-free delivery over a lossy channel. As we have already discussed in Chapter 9, the underlying assumptions do not always apply but in general it is not a bad starting point.

Rate–distortion theory has conventionally been used to establish performance bounds (in terms of bit rate and quality) for given types of source, usually characterized in terms of their signal statistics. One of the problems faced in this respect is that simple statistical characterizations do not usually fit well with complex non-stationary sources like video.

Even in the simple case of optimizing a scalar quantizer, we still need to select an appropriate statistical model for the source. Only then can we investigate practical algorithms and assess how closely they perform to the bounds for that distribution. The problem with this approach is twofold:

1. If the model is poor then the bounds are meaningless and practical or heuristic designs may well outperform them. Ortega and Ramchandran [43] provide an example where a low complexity wavelet image encoder outperforms an encoder based on an i.i.d. Gaussian model by over 3 dB.
2. Even if the model is good, a practical solution may be unrealistically complex or require huge latency. For the example given in Ref. [43], the wavelet coder is relatively low complexity whereas the "R-D optimum" encoder is of infinite complexity.

Distortion measures

Let the distortion between a symbol of a coded source $\tilde{s}$ and its original version s be denoted as $d(s, \tilde{s})$. In the case where $s = \tilde{s}$ then $d(s, \tilde{s}) = 0$. The average distortion

across the source is given by:

$$D = d\left(\mathbf{s}, \tilde{\mathbf{s}}\right) = E\left\{d\left(s, \tilde{s}\right)\right\} = \sum_{s}\sum_{\tilde{s}} P\left(s, \tilde{s}\right) d\left(s, \tilde{s}\right) \tag{10.14}$$

Normally the distortion measure used is either squared error, $d(s, \tilde{s}) = (s - \tilde{s})^2$, or absolute error, $d(s, \tilde{s}) = |s - \tilde{s}|$. The average distortion for a source over N samples, for the case of a squared error distortion, is the MSE:

$$d\left(\mathbf{s}, \tilde{\mathbf{s}}\right) = \frac{1}{N}\sum_{n=1}^{N} d_n\left(s, \tilde{s}\right) = \frac{1}{N}\sum_{n=1}^{N}\left(s[n] - \tilde{s}[n]\right)^2 \tag{10.15}$$

More often, in practical RDO applications the Sum of Squared Differences (SSD) is used, which is computed as in equation (10.15), but omitting the $1/N$ term.

The memoryless Gaussian source

As we have discussed above, closed form solutions for R-D bounds are in general difficult to obtain. However, for the case of a Gaussian i.i.d. source with variance σ^2, the bound can be computed (assuming high bit rates) and is given by:

$$\bar{R}(D) = \begin{cases} \frac{1}{2}\log_2\left(\frac{\sigma^2}{D}\right) & 0 \le D < \sigma^2 \\ 0 & D \ge \sigma^2 \end{cases} \tag{10.16}$$

where $R(D)$ denotes the rate–distortion function (which gives the bit rate R needed to deliver a distortion D for a given source) and $\bar{R}(D)$ is the rate–distortion bound (i.e. the minimum rate among all possible coders for an infinite length vector).

In practice, for more general and complex sources such as natural video, theoretical R-D bounds serve little purpose and more pragmatic approaches must be adopted. These are considered next.

10.5.2 Practical rate–distortion optimization

Operational control of a source encoder is a major issue in video compression and the aim of practical RDO is to select the best coding modes and parameter settings for the prevailing operating conditions. The available bit budget must thus be allocated in a manner such that the overall distortion is kept as low as possible in accordance with a given rate constraint. Parameter settings for modern codecs will cover a wide range of encoder attributes, such as block size, motion estimation references, inter- or intra-mode selection, and quantization step size. The optimization problem is made particularly difficult due to the non-stationary nature of video content, where the best encoder settings can vary significantly for different spatio-temporal regions. Furthermore the rate–distortion costs are not generally independent for all coding units (for example, due to the effects of prediction).

It should be noted that, especially for lower bit rates, it is not practical to consider quantization and entropy coding independently. Unlike higher bit rates and quality settings, where the number of alphabet symbols is large, for the case of lower bit rates it may for example be advantageous to choose a quantization point that delivers slightly higher distortion because it leads to significantly lower bit rate.

From source statistics to a parameterizable codec

Practical rate–distortion optimization methods do not normally start with a source model. Instead they are based on the assumption of a given parameterizable encode–decode combination which is known to perform well if the right parameters are chosen. In such cases, parameter choices are not based on source statistics but rather on some transformation of the source data (e.g. DCT coefficients or wavelet subband coefficients). This has the benefit of making the modeling more robust as the statistics of transformed data tend to be easier to model. For example, subband coefficients can be roughly approximated to a memoryless source modeled with an i.i.d. Laplacian process.

However, if the subband coefficients were indeed memoryless, we could entropy encode them based on this assumption using first order entropy. In practice the assumption is not sufficiently good as the coefficients, when scanned after quantization, typically present long chains of zeros. Hence the need for tricks such as zig-zag run-value coding that we explored in Chapter 7.

RDO methods must find the best operating points based on the optimum parameter vector, $\mathbf{p}_{\text{opt}}$. If we evaluate all parameter combinations for a given realization of a source (e.g. an image) over a range of rates then we can obtain the best rate–distortion curve for that coder–sequence combination. The convex hull, as shown in Figure 10.9, represents the operational bound against which any practical RDO heuristic algorithm can be measured.

RDO complexity

It is worth quickly mentioning the complexity of the RDO process. In off-line cases, where a single encode will be re-used multiple times, the value proposition will justify more complex or multipass encoding (a good example is encoding for a Blu-ray disk), while in other cases it may not.

Lagrangian optimization

Lagrangian methods have become widely accepted as the preferred approach for RDO in recent standards, primarily due to their effectiveness and simplicity [46,48]. Initially we will consider the case where each coding unit, i (the basic optimization building block), can be optimized independently of all others and where the only optimization parameter is a quantization index, j. While this assumption clearly breaks down in cases where context dependence or prediction are invoked, it can provide a useful and tractable solution.

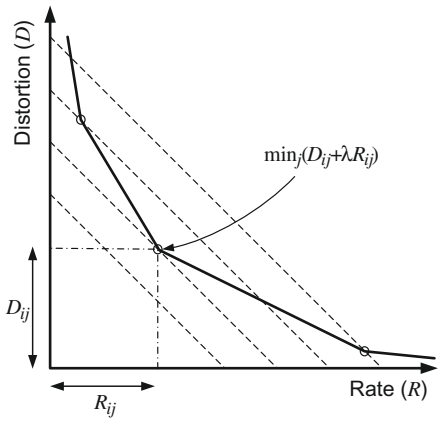


FIGURE 10.10

Lagrangian optimization for a coding unit, *i*.

Normally a discrete Lagrangian approach is adopted for RDO, as introduced by Shoham and Gersho [43,49], in order to find the maxima or minima of a function, subject to external constraints, where a closed form solution does not exist. The Lagrangian cost function for a coding unit i is given by equation (10.17):

$$J_{ij}\ (\lambda) = D_{ij} + \lambda R_{ij} \tag{10.17}$$

where the quantization index j dictates the trade-off between rate and distortion and the Lagrange multiplier λ controls the slope of lines in the R-D plane that intersects the R-D characteristic to select specific operating points. This is shown graphically in Figure 10.10.

If we minimize equation (10.17) then we obtain:

$$\frac{\partial(J_{ij}(\lambda))}{\partial R_{ij}} = \frac{\partial(D_{ij} + \lambda R_{ij})}{\partial R_{ij}} = \frac{\partial D_{ij}}{\partial R_{ij}} + \lambda_{\text{opt}} = 0 \tag{10.18}$$

This implies that the optimum value of λ is given by the negative slope of the distortion function:

$$\lambda_{\text{opt}} = -\frac{\partial D_{ij}}{\partial R_{ij}} \tag{10.19}$$

Clearly if we minimize J for the case of $\lambda = 0$ we are minimizing the distortion and if we do the same when $\lambda = \infty$ we minimize the rate. However, in general, the slope is not usually known so we must estimate it using operational techniques as described below. It should be noted that Lagrangian approaches are constrained to select operating points on the convex hull of the R-D characteristic. In some cases, preferable non-optimal points may exist. This can be addressed using alternative but potentially more complex techniques such as dynamic programming [43].

10.5.3 The influence of additional coding modes and parameters

Modern coding standards offer high coding efficiency, not just because of their underlying architecture, but also because of their adaptability to different content types and channel constraints. We will examine this in more detail in Chapter 12, but the adaptation is typically based on some or all of the following decisions:

- Quantizer control.
- When to use block or picture skip modes.
- Determination of motion vectors and multiple reference frame indices.
- The use of integer-, half- or quarter-pixel motion estimation.
- The determination of intra or inter modes and the associated block partitions. For example, in H.264/AVC the standard uses a 16×16 luma macroblock with the following options:
  - Intra modes (for all frames): nine 4×4 modes and four 16×16 modes.
  - Inter modes (only for P- and B-frames): macroblock partitions of 16×16, 16×8, 8×16, 8×8, 8×4, 4×8, 4×4.

Each mode has an associated rate–distortion cost and the encoder must select the mode having the lowest cost. This is of course complicated by the fact that the optional modes will have different efficiencies for different content types and at different bit rates. The optimization problem is complicated even further because the rate–distortion costs are not generally independent for all coding units; for example, because of the spatial and/or temporal dependencies introduced through intra- and inter-prediction. This means that the Lagrangian optimization of equation (10.17) should theoretically be performed jointly over all coding units in the video sequence, something that in most cases would be prohibitively complex. A number of simplification techniques have therefore been proposed.

Perhaps the most obvious method of determining parameter choices or mode decisions is to use an exhaustive search. By comparing all possible options and parameter combinations, the optimum can be selected. This can be prohibitively expensive, so alternatives are generally used. Much simpler strategies for decision making can be made based on indicative thresholds. For example, the decision as to whether an intra or inter mode should be selected could be based on a comparison between the SAD of a 16×16 macroblock with respect to its mean, and the minimum SAD after integer motion search. Similar strategies have been reported for other decisions. However, the primary method of choice for RDO is based on the use of Lagrange multiplier methods and these are normally applied separately to the problems of mode selection and motion estimation.

Lagrangian multipliers revisited

Following the approach above and as described in Ref. [46], the aim is to solve the constrained problem:

$$\min_{\mathbf{p}} D(\mathbf{p}) \text{ s.t. } R(\mathbf{p}) \leq R_T \tag{10.20}$$

where R_T is the target bit rate and $\mathbf{p}$ is the vector of coding parameters. As before, this can be represented as an unconstrained Lagrangian formulation based on the following cost function:

$$\mathbf{p}_{\text{opt}} = \arg\min_{\mathbf{p}} \{D(\mathbf{p}) + \lambda R(\mathbf{p})\} \tag{10.21}$$

Intra modes: Let us first consider the problem of mode selection for *intra* coding modes. If we assume that the quantizer value Q is known and that the Lagrange parameter λ_{MODE} is given, then the parameter (or coding mode) selection is given by:

$$\begin{aligned} &J_{\text{MODE}}(\mathbf{S}_k, \mathbf{p}_k \mid Q, \lambda_{\text{MODE}}) \\ &\quad = D(\mathbf{S}_k, \mathbf{p}_k \mid Q, \lambda_{\text{MODE}}) + \lambda_{\text{MODE}} \cdot R(\mathbf{S}_k, \mathbf{p}_k \mid Q, \lambda_{\text{MODE}}) \end{aligned} \tag{10.22}$$

where the parameter vector $\mathbf{p}$ is varied over all possible coding modes for the given coding unit (or sub-image), $\mathbf{S}_k$. As discussed previously, the distortion measure D is normally based on sum-of-squared differences between the original block and its reconstructed version and the rate R is normally measured after entropy coding.

Skip modes: For the case of computing *skip* modes, the distortion and rate calculations do not depend on the current quantizer value, but simply on the SSD between the current coding unit and that produced by prediction using the inferred motion vector. The rate for a skipped block is approximately 1 bit per coding unit in H.264.

Inter modes: For the case of *inter* modes, let us assume we know the Lagrange parameter λ_{MOTION} and the decoded reference picture $\tilde{\mathbf{S}}$. Rate constrained motion estimation can be performed by minimizing the cost function as follows:

$$\mathbf{p}_{\text{opt}} = \arg\min_{\mathbf{p}} \{D_{\text{DFD}}(\mathbf{S}_k, \mathbf{p}_k) + \lambda_{\text{MOTION}} \cdot R_{\text{MOTION}}(\mathbf{S}_k, \mathbf{p}_k)\} \tag{10.23}$$

The distortion measure used in this case is normally either the SAD or the SSD, as defined by equation (10.24) with $l = 1$ or $l = 2$ respectively:

$$D_{\text{DFD}}(\mathbf{S}_k, \mathbf{p}_k) = \sum \left| s\left[x, y, z\right] - \tilde{s}[x - d_x, y - d_y, z - d_z] \right|^l \tag{10.24}$$

where $\mathbf{d}$ is the motion vector for the current parameter set. The rate calculation for motion must include all the bits required to code the motion information, normally based on entropy coded predictions.

Wiegand et al. [44,46] demonstrated the power of the Lagrangian multiplier approach in the context of block size selection, motion estimation, and quantizer control. In the context of quantizer control, they demonstrated the relationship between quantizer step size and λ. They showed that, to a first order approximation, this relationship is content independent and (in the context of H.263) given by:

$$\lambda_{\text{MODE}} = 0.85 Q^2_{\text{H.263}} \tag{10.25}$$

The Lagrange approach has been used widely since its introduction in H.263 and has been adopted in the test models for H.264/AVC and HEVC. For the case of

H.264/AVC, a different expression has been determined empirically to give for I and P frames:

$$\lambda_{\text{MODE(I,P)}} = 0.85 \cdot 2^{(\text{QP}_{\text{H.264}}-12)/3} \tag{10.26}$$

and for B frames:

$$\lambda_{\text{MODE(B)}} = \max\left(2, \min\left(4, \frac{\text{QP}-12}{6}\right)\right) \cdot \lambda_{\text{MODE(I,P)}} \tag{10.27}$$

The relationship between λ_{MODE} and λ_{MOTION} is given by:

$$\begin{cases} \lambda_{\text{MOTION}} = \lambda_{\text{MODE}}: & \text{for SSD} \\ \lambda_{\text{MOTION}} = \sqrt{\lambda_{\text{MODE}}}: & \text{for SAD} \end{cases} \tag{10.28}$$

10.5.4 From Rate–Distortion to Rate–Quality Optimization

The techniques described throughout this section have been based on two assumptions: (i) that the aim of the encoding process is to represent the picture sample values as closely as possible, and (ii) that MSE (or other L_p norms) accurately reflects the perceptual quality and importance of the spatio-temporal region being coded. In practice we have seen that, although the second assumption is clearly violated, MSE can provide monotonic and consistent indicators for parameter variations within a given coding strategy for given content [20].

In the context of the first assumption, new perceptually inspired coding strategies are emerging based on texture analysis and synthesis models (see Chapter 13). The aim of these is not to minimize the distance between the original and coded version, but instead to obtain a perceptually plausible solution. In this case, MSE is unlikely to provide any meaningful indication of perceived quality and will no longer be a valid objective function. Emphasis must therefore shift from Rate–Distortion Optimization to Rate–Quality Optimization, demanding new embedded perceptually driven yet low complexity quality metrics [23].

10.6 Rate control

Rate control is an important component in a practical video codec since it ensures that the coded video is delivered to the channel at a rate compatible with the prevailing content complexity and channel capacity. Putting to one side for the moment the influences of specific coding modes, the primary issues that influence the bit rate of an encoded video are:

1. The different requirements for I-pictures and P- or B-pictures (I-pictures can consume 5–10 times the number of bits of a P- or B-picture).
2. The influence of source complexity. Content with high spatio-temporal activity will normally require many more bits than for the case of low activity content. This is depicted in Figure 10.11.

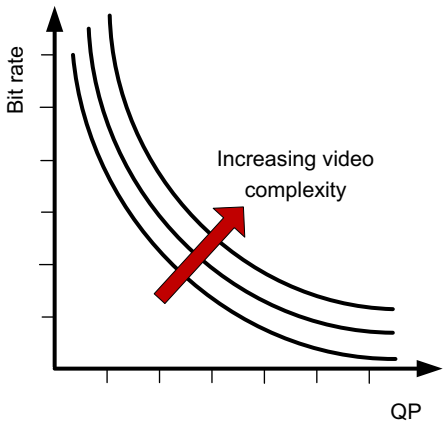


FIGURE 10.11

Managing bit rate in the context of changing source complexity.

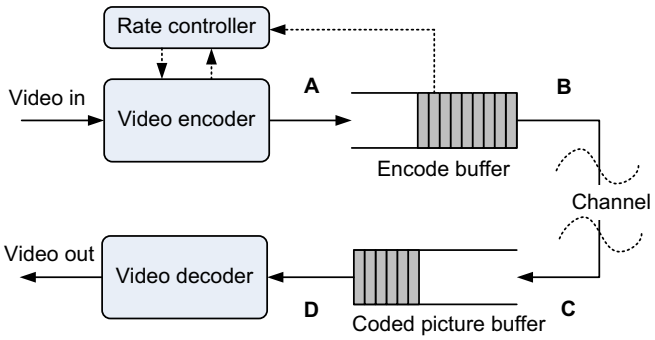


FIGURE 10.12

Transmission buffering at encoder and decoder.

Hence if the quantizer is kept fixed, then the peak-to-mean ratio of the output bitstream could be excessively high, translating to an efficient use of bandwidth. The peak output bit rate could also exceed the maximum channel capacity.

Even when rate control is used, bit rate fluctuations are inevitable and buffering is invariably employed to smooth out these variations. A simple depiction of the interface between an encoder and the channel and the channel and a decoder via buffers is shown in Figure 10.12. This figure explicitly includes a rate controller that is used to moderate the flow of coded data into the encode buffer. This is normally achieved by means of a rate-quantization model that delivers a quantizer parameter (QP) to the encoder based on dynamic measures of buffer occupancy and content complexity. If we consider the delays due to the buffering operations and the channel, we can represent the playout schedule for this type of configuration in Figure 10.13.

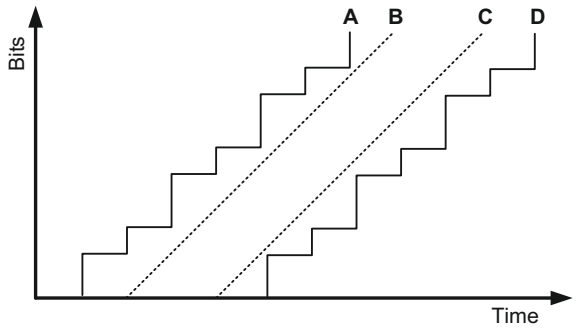


FIGURE 10.13
Streaming playout schedule on a picture by picture basis, corresponding to the points labeled in Figure 10.12.

The buffers described above and the associated rate control mechanisms are non-normative in most standards. However, to ensure that rate control is performed correctly at the encoder, recent standards have incorporated the concept of a *Hypothetical Reference Decoder* (HRD). This is used to emulate idealized decoder behavior in terms of buffering and bit consumption. Without rate control, decoder buffers would regularly underflow or overflow, resulting in playout jitter or loss.

Rate control and RDO jointly present a dilemma: in order to perform RDO for a coding unit, the quantization parameter must be known, and this is usually based on SAD or a variance measure. However, the actual SAD of the coding unit is not available until after the RDO process is complete; it is thus necessary to estimate the SAD of the current coding unit. Alongside this we also need to compute target bits for the current frame, and this is complicated by the fact that header information such as coding unit modes and MV information are not available before the RDO process completes.

10.6.1 Buffering and HRD

Bit rate allocation is normally associated with a buffer model that is specified as part of a Hypothetical Reference Decoder (HRD) in the video coding standard. The HRD is usually a normative component of a standard since it is important in dictating the requirements on compliant bitstreams.

The HRD is an idealized representation of a decoder that is conceptually connected to the output of the encoder. It comprises a decoder buffer as well as the decoder itself. The buffer is normally modeled using a leaky bucket approach which simply provides a constant rate flow of bits to and from the channel. Bits are assumed to flow into the decoder buffer (sometimes called the Coded Picture Buffer (CPB)) at a constant rate and are assumed to be extracted from it by the decoder instantaneously in picture-sized blocks. This is the scenario depicted in Figures 10.12 and 10.13 where, in the absence of loss in the channel, the content of the buffer at the receiver simply mirrors

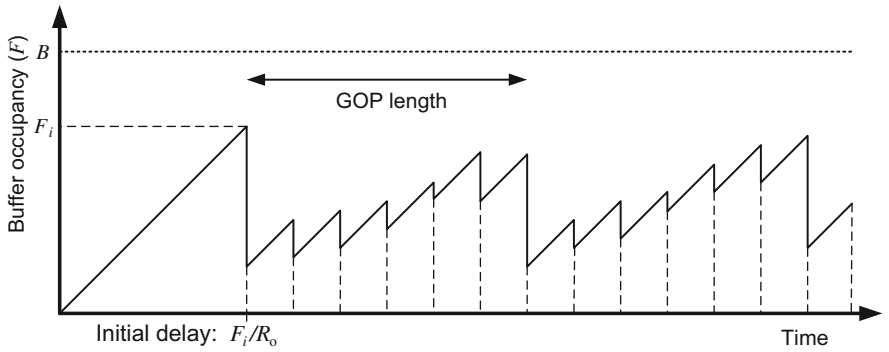


FIGURE 10.14

Example decoder buffer timing, indicating HRD parameters.

that at the encoder. An HRD-compliant bit stream must thus be processed by the CPB without overflow and underflow. This requirement must be satisfied by the rate control implemented in the encoder.

The HRD buffer model is normally described by three parameters, (i) the output rate R_o (assumed constant), the initial occupancy, F_i (which is the occupancy level at which decoding is assumed to start), and the buffer size, B. The rate controller must track the occupancy F of the hypothetical buffer and adjust the encoder parameters within its scope, in order to avoid buffer underflow and overflow (see Figure 10.14).

Example 10.2 (Buffering and HRD parameters)

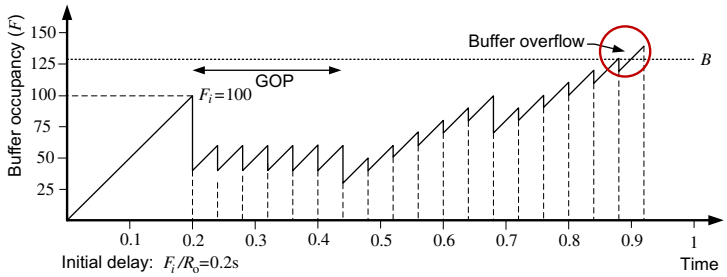
Consider the operation of a constant bit rate video transmission system at 24 fps with the following parameters:

$$R_o = 500 \text{ kbps};\ \ B = 128 \text{ kb};\ \ F_i = 100 \text{ kb};\ \ \text{GOP} = 6 \text{ frames}$$

If the initial pictures transmitted have the following sizes, compute the occupancy of the decoder buffer and determine whether underflow or overflow occurs.

Frame no.	Picture size (kbits)
1	60
2–6	20
7	30
8–12	10
13	30
14–18	10
19	50
20–24	30

Solution. The ramp-up time for the buffer is $F_i/R_o = 0.2$ s. Plotting the buffer occupancy over the 1 s period of these initial pictures, we can observe that the decoder buffer overflows when pictures 18 and 19 are transmitted:



10.6.2 Rate control in practice

Figure 10.15 shows a generic rate control architecture. While simplified, this has many similarities to those used in MPEG-2, H.264/AVC, and HEVC. The basic elements are described below.

Buffer model

The buffer model represents the hypothetical reference decoder as described in Section 10.6.1. It is parameterized by its size and initial fullness, and takes a further input according to the bit rate of the channel. It provides an output of its current occupancy, F, to the bit allocation blocks described below.

Complexity estimation

In order to control the bit allocation process it is essential to estimate the complexity of the coded residual. The estimation of source complexity is normally based on the residual data (since this is what the quantizer influences). A measure such as Mean Absolute Difference (MAD), Sum of Absolute Values (SAD), Sum Squared Difference (SSD), or Sum of Absolute Transform Differences (SATD) [50] of the residual is often used for this purpose.

Rate-quantization model

Reflecting the variability indicated in Figure 10.11, the rate-quantization model describes how the available bits are related to quantizer values when content of varying complexity is encoded.

ΔQP limiter

In order to avoid excessive changes in QP values that might introduce noticeable artifacts or cause oscillatory behavior in the rate controller, the change in QP value is normally limited to a small increment, typically ± 2.

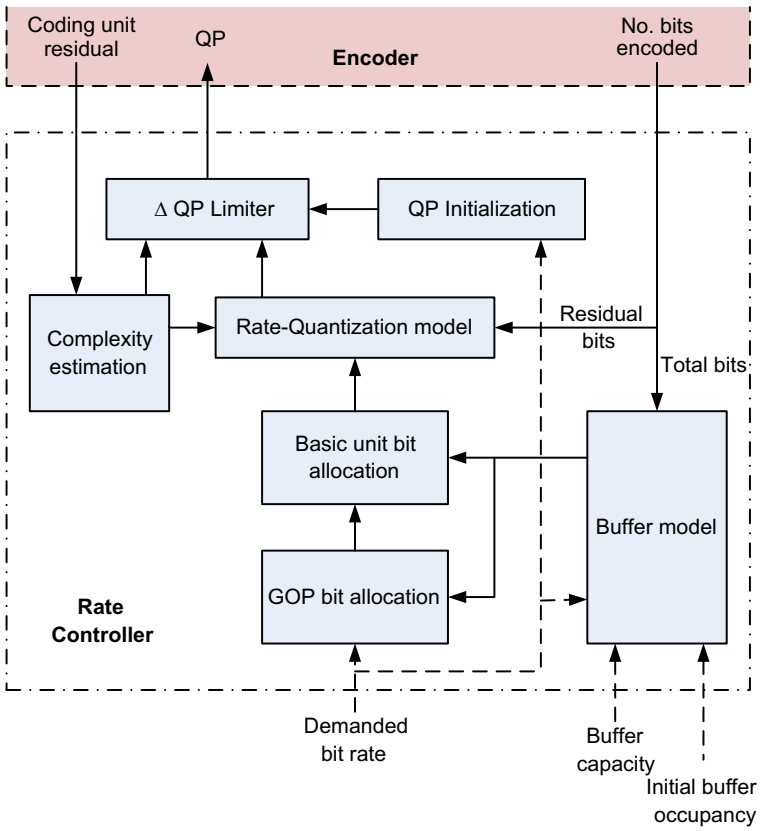


FIGURE 10.15

Generic rate controller.

QP initialization

The QP value must be set to some value prior to coding. An average value for the bit rate per pixel can be easily computed, based on the bit rate demanded and the video format, thus:

$$\text{bpp} = \frac{R_1(1)}{f \cdot X \cdot Y} \tag{10.29}$$

where $R_i(k)$ is the instantaneous available bit rate at the time of coding frame k of GOP i, f is the frame rate, and X and Y are the number of horizontal and vertical pixels in the picture. This can then be related to QP values by way of a look-up table such as that recommended for H.264/AVC in Ref. [51].

GOP bit allocation

Based on the demanded bit rate and the buffer occupancy, a target bit allocation for the GOP is computed. GOP level bit allocation normally initializes the IDR-picture and the first P-picture. A scheme used for this in H.264/AVC is as follows [52]:

$$\mathrm{QP}_i(1) = \max\left\{\mathrm{QP}_{i-1}(1) - 2, \min\left\{\mathrm{QP}_{i-1}(1) + 2, \frac{\sum_{k=2}^{N_{i-1}} \mathrm{QP}_{i-1}(k)}{N_{i-1} - 1}, - \min\left\{2, \frac{N_{i-1}}{15}\right\}\right\}\right\} \tag{10.30}$$

where $\mathrm{QP}_i(k)$ is the QP value assigned to picture k of GOP i and N_i is the total number of pictures in GOP i. If $\mathrm{QP}_i(1) > \mathrm{QP}_{i-1}(N_{i-1}) - 2$, this is further adjusted as follows:

$$\mathrm{QP}_i(1) \leftarrow \mathrm{QP}_i(1) - 1 \tag{10.31}$$

where $\mathrm{QP}_{i-1}(N_{i-1})$ represents the QP value of the previous P-picture.

Coding unit bit allocation

This refers to the process of allocating bits to smaller coding elements, such as individual pictures, slices, or smaller coding units. The bit allocation to each picture is based on the target buffer level, actual buffer occupancy, the bit rate demanded, and of course the frame rate. QP values for each of the remaining P- and B-frames are thus assigned based on the number of remaining bits available for the GOP. The allocation process should also take account of whether the current picture is of P or B type since these may be provided with different bit allocations. The scheme used for this in H.264/AVC, based on a quadratic model derived assuming a Laplacian source distribution, is as follows [52]:

Firstly, the MAD of the current stored picture is estimated, $\tilde{\xi}_i(k)$, based on the corresponding value for the previous picture, as given in equation (10.32):

$$\tilde{\xi}_i(k) = a_1 \cdot \xi_i(k-1) + a_2 \tag{10.32}$$

where a_1 and a_2 are the coefficients of the prediction model, initialized to 1 and 0 and updated after each unit is coded. Next, the quantizer step size is computed for each of the remaining P-pictures, using equation (10.33).

$$T_i(k) = c_1 \cdot \frac{\tilde{\xi}_i(k)}{Q_{\mathrm{step},i}(k)} + c_2 \cdot \frac{\tilde{\xi}_i(k)}{Q^2_{\mathrm{step},i}(k)} + m_{h,i}(k) \tag{10.33}$$

where $m_{h,i}(k)$ is the total number of header bits and motion vector bits, c_1 and c_2 are the coefficients of the quadratic model, updated after each picture is processed, and $Q_{\mathrm{step},i}(k)$ is the quantizer step size for picture k, corresponding to the target number

of bits. In order to compute the quantizer step size in equation (10.33) we first need to calculate a value for the target number of bits for frame k, $T_i(k)$. This is usually computed based on a weighted combination of the buffer occupancy and the number of bits remaining for the rest of the GOP. Further details on this scheme can be found in Ref. [52].

Several other schemes have been proposed for rate control that improve prediction accuracy for both texture and header bits; for example, the contribution of Kwon et al. [50]. In the context of HEVC, several modifications to the above approach have been proposed and the reader is referred to the contributions by Sanz-Rodriguez and Schierl [53] and by Li et al. [54].

10.6.3 Regions of interest and rate control

In certain applications, and for certain types of content, a priori information about the scene, or areas within it, can be exploited to improve coding efficiency or better manage bit rate. For example, in many surveillance videos, there may be one or more *regions of interest* (ROI) that would benefit from prioritized bit allocation. A good discussion of the trade-offs involved in this is provided by Sadka [55].

Agrafiotis et al. [56] presented a modified rate control method that allows the definition of multiple priority regions, the quality of which varies, based on the ROI characteristics and the target bit rate. The results presented demonstrate increased coding flexibility which can lead to significant quality improvements for the ROI and a perceptually pleasing variation in quality for the rest of the frame, without violating the target bit rate. This work was presented in the context of low bit rate sign language coding for mobile terminals, but it can also be of benefit in many other applications including surveillance and healthcare.

10.7 Summary

This chapter has covered a range of methods for measuring and controlling visual quality. We first described the most common subjective testing methods and showed how these can provide a consistent means of comparing codec performance. We then went on to discuss a number of objective measures that can be used as an efficient substitute for subjective evaluation in certain cases. MSE-based methods were reviewed briefly and their strengths and weaknesses highlighted. We demonstrated cases where the perceptual distortion experienced by the human viewer cannot however be characterized using such simple mathematical differences. Perception-based video metrics have therefore emerged and a number of these were reviewed and compared, showing improved correlation with subjective scores.

The second part of this chapter addressed the issue of rate–distortion optimization, describing some of the most common techniques that enable us to select the optimum coding parameters for each spatio-temporal region of a video. After reviewing conventional rate–distortion theory, we focused on practical solutions such as those employing Lagrange multipliers, showing that these can provide a tractable

solution for modern codecs. Related to this, the final part of the chapter studied the requirements of, and methods for, rate control as a means of adapting the output bit rate of an encoder according to the capacity of a given channel.

References

[1] S. Winkler, Digital Video Quality: Vision Models and Metrics, Wiley, 2005.

[2] F. Kozamernik, P. Sunna, E. Wycken, D. Pettersen, Subjective quality assessment of internet video codecs — phase 2 evaluations using SAMVIQ, EBU Technical Review (2005) 1–22.

[3] S. Winkler, Analysis of public image and video database for quality assessment, IEEE Journal of Selected Topics in Signal Processing 6 (6) (2012) 1–10.

[4] Recommendation ITU-R BT.500-13, Methodology for the Subjective Assessment of the Quality of Television Pictures, ITU-R, 2012.

[5] Recommendation ITU-T P.910, Subjective Video Quality Assessment Methods for Multimedia Applications, ITU-T, 1999.

[6] F. De Simone, L. Goldmann, J.-S. Lee, T. Ebrahomi, Towards high efficiency video coding: subjective evaluation of potential coding methodologies, Journal of Visual Communication and Image Representation 22 (2011) 734–748.

[7] P. Hanhart, M. Rerabek, F. De Simone, T. Ebrahimi, Subjective quality evaluation of the upcoming HEVC video compression standard, in: Applications of Digital Image Processing XXXV, vol. 8499, 2012.

[8] H. Hoffmann, T. Itagaki, D. Wood, T. Hinz, T. Wiegand, A novel method for subjective picture quality assessment and further studies of HDTV formats, IEEE Transactions on Broadcasting 54 (1) (2008) 1–13.

[9] D. Lane, Introduction to Statistics. <http://onlinestatbook.com/>, Rice University.

[10] Video Quality Experts Group, Final Report from the Video Quality Experts Group on the Validation of Objective Quality Metrics for Video Quality Assessment, VQEG, Technical Report, 2000. <http://www.its.bldrdoc.gov/vqeg/projects/frtv-phase-i/frtv-phase-i.aspx> phase I.

[11] Video Quality Experts Group, Final VQEG Report on the Validation of Objective Models of Video Quality Assessment, VQEG, Technical Report, 2003. <http://www.its.bldrdoc.gov/vqeg/projects/frtv> phase II.

[12] H. Sheikh, Z. Wang, L. Cormack, A. Bovik, LIVE Image Quality Assessment Database Release 2. <http://live.ece.utexas.edu/research/quality>.

[13] K. Seshadrinathan, A. Bovik, Motion tuned spatio-temporal quality assessment of natural videos, IEEE Transactions on Image Processing 19 (2010) 335–350.

[14] K. Seshadrinathan, R. Soundararajan, A.C. Bovik, L.K. Cormack, Study of subjective and objective quality assessment of video, IEEE Transactions on Image Processing 19 (2010) 335–350.

[15] P. Le Callet, F. Autrusseau, Subjective quality assessment IRCCyN/IVC database, 2005. <http://www.irccyn.ec-nantes.fr/ivcdb/>.

[16] Video Quality Experts Group, Report on the Validation of Video Quality Models for High Definition Video Content, VQEG, Technical Report, 2010. <http://www.its.bldrdoc.gov/vqeg/projects/hdtv/hdtv.aspx>.

[17] Video Quality Experts Group, Final Report from the Video Quality Experts Group on the Validation of Objective Models of Multimedia Quality Assessment, VQEG,

Technical Report, 2008. <http://www.its.bldrdoc.gov/vqeg/projects/multimedia-phase-i/multimedia-phase-i.aspx>.

[18] P. Marziliano, F. Dufaux, S. Winkler, T. Ebrahimi, A no-reference perceptual blur metric, in: Proceedings of the IEEE International Conference on Image Processing, vol. 3, 2002, pp. 57–60.

[19] R. Soundararajan, A. Bovic, Video quality assessment by reduced reference spatio-temporal entropic differencing, IEEE Transactions on Circuits and Systems for Video Technology 23 (4) (2013) 684–694.

[20] D. Huynh-Thu, M. Ghanbari, Scope of validity of PSNR in image/video quality assessment, Electronics Letters 44 (13) (2008) 800–801.

[21] Z. Wang, A. Bovik, Mean squared error: love it or leave it? IEEE Signal Processing Magazine 26 (1) (2009) 98–117.

[22] B. Girod, What's wrong with mean squared error? in: A. Watson (Ed.), Digital Images and Human Vision, MIT Press, 1998.

[23] F. Zhang, D. Bull, A parametric framework for video compression using region-based texture models, IEEE Journal of Selected Topics in Signal Processing 5 (7) (2011) 1378–1392.

[24] T. Pappas, T. Michel, R. Hinds, Supra-threshold perceptual image coding, in: Proceedings of the IEEE International Conference on Image Processing, 1996, pp. 234–240.

[25] V. Kayargadde, J. Martens, Perceptual characterization of images degraded by blur and noise: experiments, Journal of the Optical Society of America 13 (1996) 1166–1177.

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

[27] M. Carnec, P. LeCallet, D. Barba, An image quality assessment method based on perception of structural information, in: Proceedings of the IEEE International Conference on Image Processing, vol. 2, 2003, pp. 185–188.

[28] D. Chandler, S. Hemami, VSNR: a wavelet-based visual signal to noise ratio for natural images, IEEE Transactions on Image Processing 16 (9) (2007) 2284–2298.

[29] E. Larson, D. Chandler, Most apparent distortion: full reference image quality assessment and the role of strategy, Journal of Electronic Imaging 19 (1) (2010), pp. 011006-1–011006-21.

[30] S. Chikkerur, V. Sundaram, M. Reisslein, L. Karam, Objective video quality assessment methods: a classification, review and performance comparison, IEEE Transactions on Broadcasting 30 (2005) 17–26.

[31] W. Lin, C. Kuo, Perceptual visual quality metrics: a survey, Journal of Visual Communication and Image Representation 22 (2011) 297–312.

[32] S. Sheikh, A. Bovik, Image information and visual quality, IEEE Transactions on Image Processing 15 (2006) 430–444.

[33] S. Winkler, P. Mohandas, The evolution of video quality measurement: from PSNR to hybrid metrics, IEEE Transactions on Broadcasting 54 (3) (2008) 660–668.

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

[35] Z. Wang, A. Bovik, H. Sheikh, E. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Transactions on Image Processing 13 (2004) 600–612.

[36] Z. Wang, E. Simoncelli, Translation insensitive image similarity in complex wavelet domain, in: IEEE International Conference on Acoustics, Speech and Signal Processing, 2005, pp. 573–576.

[37] Z. Wang, E. Simoncelli, A. Bovik, Multi-scale structural similarity for image quality assessment, in: Proceedings of the IEEE Asilomar Conference on Signals, Systems and Computers, 2003, pp. 1398–1402.
[38] Z. Wang, L. Lu, A. Bovik, Video quality assessment based on structural distortion measurement, Signal Processing: Image Communication 19 (2) (2004) 121–132.
[39] P. Vu, C. Vu, D. Chandler, A spatiotemporal most apparent distortion model for video quality assessment, in: Proceedings of the IEEE International Conference on Image Processing, 2011, pp. 2505–2508.
[40] R. Mantiuk, K. Kim, A. Rempel, W. Heidrich, HDR-VDP-2: a calibrated visual metric for visibility and quality predictions in all luminance conditions, ACM Transactions on Graphics (Proceedings of SIGGRAPH'11) 30 (4) (2011) (article no. 40).
[41] A. Zhang, D. Bull, Quality assessment method for perceptual video compression, in: Proceedings of the IEEE International Conference on Image Processing, 2013, pp. 39–43.
[42] A. Zhang, D. Bull, A perception-based hybrid model for video quality assessment, IEEE Transactions on Circuits and Systems for Video Technology, submitted for publication.
[43] A. Ortega, K. Ramchandran, Rate-distortion methods for image and video compression, IEEE Signal Processing Magazine 15 (6) (1998) 23–50.
[44] G. Sullivan, T. Wiegand, Rate-distortion optimization for video compression, IEEE Signal Processing Magazine 15 (6) (1998) 74–90.
[45] Y. Wang, J. Ostermann, Y. Q Zhang, Video Processing and Communications, Prentice Hall (2002).
[46] T. Wiegand, H. Schwarz, A. Joch, F. Kossentini, G. Sullivan, Rate-constrained coder control and comparison of video coding standards, IEEE Transactions on Circuits and Systems for Video Technology 13 (7) (2003) 688–703.
[47] E.H. Yang, X. Yu, Rate distortion optimization for H.264 interframe coding: a general framework, IEEE Transactions on Image Processing 16 (7) (2007) 1774–1784.
[48] H. Everett, Generalised Lagrange multiplier method for solving problems of optimum allocation of resources, Operations Research 11 (1963) 399–417.
[49] Y. Shoham, A. Gersho, Efficient bit allocation for an arbitrary set of quantizers, IEEE Transactions on Acoustics, Speech, and Signal Processing 36 (1988) 1445–1453.
[50] D.-K. Kwon, M.-Y. Shen, C. Kuo, Rate control for H.264 video with enhanced rate and distortion models, IEEE Transactions on Circuits and Systems for Video Technology 17 (5) (2007) 517–529.
[51] G. Sullivan, T. Wiegand, K. Lim, Joint Model Reference Encoding Methods and Decoding Concealment Methods, JVT-I049, San Diego, 2003.
[52] K. Keng-Pang Lim, G. Sullivan, T. Wiegand, Joint Video Team (JVT) of ISO/IEC MPEG and ITU-T VCEG Document JVT-O079, Text Description of Joint Model Reference Encoding Methods and Decoding Concealment Methods, 2005.
[53] S.S. Sanz-Rodrıguez, T. Schierl, A rate control algorithm for HEVC with hierarchical GOP structures, in: Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing, 2013, pp. 1719–1723.
[54] B. Li, H. Li, L. Li, J. Zhang, Rate control by R-lambda model for HEVC, JCTVC-K0103, in: 11th JCTVC Meeting, China, October 2012.
[55] A. Sadka, Compressed Video Communications, Wiley, 2002.
[56] D. Agrafiotis, D. Bull, C. Canagarajah, N. Kamnoonwatana, Multiple priority region of interest coding with H.264, in: Proceedings of the IEEE International Conference on Image Processing, 2006, pp. 53–56.