INSTITUTO SUPERIOR TÉCNICO
Universidade Técnica de Lisboa

Quality Evaluation of Coded Video

Luís Miguel Malveiro Pereira Tomaz Roque

Dissertation submitted for obtaining the degree of
Master in Electrical and Computer Engineering

Jury

Supervisor:	Professora Doutora Maria Paula dos Santos Queluz
Co-Supervisor:	Engenheiro Tomás Gomes da Silva Serpa Brandão
President:	Professor Doutor José Manuel Bioucas Dias
Members:	Professor Doutor Paulo Luis Serras Lobato Correia

November 2009

Acknowledgements

First of all, I would like to thank to my supervisors Maria Paula Queluz and Tomás Gomes Brandão for the unique opportunity to perform this thesis and for the constant knowledge and experience sharing. Their orientation, availability, guidelines, advising, opinion, and constant support, were a key factor for the completion of this work, and will also be useful for my professional future.

To all my friends from Instituto Superior Técnico, for all the moments spent during the academic life, especially Luís Gomes, João Nobre, Filipe Leonardo and Inês for all the help and encouragement during the last six years and to the Algarve and Trips group for making my summers even better.

To my brother João Pedro whose advice I've always sought.

To my grandparents, who always gave me all support and love to achieve this goal.

And, finally, to my mother and father for their unconditional love and their believe that I can accomplish anything I purpose myself to do.

Abstract

With the multimedia communications emergence, there has been an increasing need to develop quality measurements techniques that can predict perceived video quality automatically. In this dissertation two different strategies for video quality measurement, in the presence of distortions due to compression, are considered. These two different video quality assessment metrics are commonly known as Subjective Quality Metrics and Objective Quality Metrics. With regard to the first one, a subjective video quality assessment test session was conducted, in order to achieve, from a number of human observers, a subjective quality measurement, the Mean Opinion Score (MOS), for a group of representative video sequences. This first method has been regarded for many years as the most reliable for quality measurement; however, this assessment method is highly time consuming and requires appropriated viewing conditions. In order to provide an automatic evaluation and monitoring of video data quality, a Mean Opinion Score prediction model based in objective quality metrics is also proposed in this dissertation. The goal of this type of video quality assessment measurement is to design an automated quality assessment method that correlates well with subjective quality assessment and, as consequence, with human visual perception. The performance of this second video quality evaluation method is validated by confronting the resulting quality measures with the scores produced by human judgment (subjective tests), and using performance metrics proposed by VQEG (Video Quality Expert Group). This second strategy provided good results being able to predict video quality scores close to those resulting from subjective assessment.

Keywords

Subjective Video Quality Metrics, Video Quality Assessment, Objective Video Quality Metrics, Mean Opinion Score.

Resumo

Com o desenvolvimento das comunicações multimédia, tem havido uma necessidade crescente de desenvolver métodos que permitam avaliar a qualidade de vídeo codificado de uma forma automática. Neste projecto são consideradas duas diferentes estratégias para medir a qualidade de video na presença de distorções devidas à compressão. Essas duas diferentes abordagens de avaliação de qualidade de vídeo são geralmente conhecidas como métricas de qualidade subjectiva e métricas de qualidade objectiva. Relativamente à primeira, foi feita uma sessão de testes de avaliação de qualidade de vídeo subjectiva, com o objectivo de obter, a partir de um determinado número de observadores humanos, medidas de qualidade subjectiva, o *Mean Opinion Score*, para um grupo representativo de sequências de vídeo. Este primeiro método tem sido considerado ao longo do tempo como o mais consistente para a medição de qualidade; contudo, este método de avaliação é demorado e requer condições de visualização apropriadas. Com o objectivo de fornecer uma avaliação automática bem como a monitorização da qualidade de vídeo, é proposto nesta dissertação um modelo de estimação do *Mean Opinion Score*, baseado em métricas de qualidade objectivas. A principal razão para desenvolver este sistema de avaliação de qualidade prende-se com o facto de se pretender um método de avaliaçao de qualidade autónomo que se correlacione bem com a avaliação de qualidade subjectiva. O desempenho deste segundo método de avaliação de qualidade de vídeo foi validado confrontando os valores resultantes deste modelo com a pontuação atribuída pelos observadores durante os testes subjectivos. Para tal, utilizaram-se métricas que permitem estabelecer uma relação quantitativa entre os métodos subjectivos e objectivos, propostas pelo VQEG (Video Quality Expert Group). Verificou-se que, de um modo geral, o método proposto produz bons resultados, sendo capaz de estimar uma pontuação de qualidade de vídeo próxima da pontuação resultante de uma avaliação subjectiva.

Palavras-chave

Métricas Subjectivas de Qualidade de Vídeo, Avaliação de Qualidade Vídeo, Métricas Objectivas de Qualidade de Vídeo, *Mean Opinion Score*.

Table of Contents

List of Figures

List of Tables

List of Acronyms

ACR	Absolute Category Rating
CIF	Common Intermediate Format
DCR	Degradation Category Rating
DCT	Discrete Cosine Transform
DSCQS	Double Stimulus Continuous Quality Scale
DSIS	Double Stimulus Impairment Scale
FR	Full Reference
HDTV	High Definition Television
HVS	Human Visual System
IEEE	Institute of Electrical and Electronics Engineers
ITU	International Telecommunications Union
MOS	Mean Opinion Score
MSE	Mean Squared Error
MSU	Moscow State University
NR	No Reference
PC	Pair Comparison Method
PCA	Principal Component Analysis
PSNR	Peak Signal-to-Noise Ratio
PVQM	Perceptual Video Quality Metric
QCIF	Quarter Common Intermediate Format
QoS	Quality of Service
RMS	Root Mean Square Error
RR	Reduced Reference
SDSCE	Simultaneous Double Stimulus for Continuous Evaluation
SIF	Source Input Format
SS	Single Stimulus Method
SSCQE	Single Stimulus Continuous Quality Evaluation
SSIM	Structural Similarity index
VQEG	Video Quality Experts Group
VQM	Video Quality Metric

Chapter 1

Introduction

The assessment of image quality in video and image processing systems plays an important role in deciding the quality of service in image and video communications, network maintenance and even to compare different service providers. Quality assessment systems have a wide range of applications from security services to entertainment, which includes digital television, internet video and in general the world of digital multimedia communications. However, the automatic evaluation of digital imaging systems quality is a challenging task since it requires either to match human perfection or to overcome human limitations. In order to give an overview of this problem difficulty, it is necessary to understand the numerous factors that contribute to what a viewer perceives as “video quality”. Among these factors are the individual interests, quality expectations, viewing conditions and display type and properties. The wide variety and subjectivity of some of these factors are indicators of the complexity of the quality measurement problem [Wink07]. The main objective of the present work is to approach the behavior of human visual system in video quality evaluation.

In order to develop and standardize the required technology for assessing the performance of digital video processing and communication systems, some organizations were formed. As example of

that, is the Institute for Telecommunication Sciences (ITS) an American organization which is responsible for the promotion of advanced telecommunications and information infrastructure development. The Institute began in the 1940s, however it was mainly from 1994 to 1997, that ITS gave a large contribution for the development of American standards for gauging the quality of digital video systems. These standards[1] were named as ANSI T1.801.01, ANSI T1.801.02, ANSI T1.801.03 and ANSI T1.801.04. Generically, the ANSI T1.801.01 provides a standardized set of video test scenes in digital format that can be used for subjective and objective testing of digital video systems, while the ANSI T1.801.02 standard provides a general description of digital video performance terms and impairments. Standard ANSI T1.801.03 defines a whole new framework of objective parameters that can be used to measure the quality of digital video systems, while the ANSI T1.801.04 standard describes metrics for audio delay, video delay, and audio-visual synchronization, since these parameters are important, particularly for interactive telecommunications services.

In October of 1997, the Video Quality Experts Group[2] (VQEG) was established in order to address video quality issues. The VQEG primary mission is to validate objective video/multimedia quality metrics and to report results to the ITU-T Study Groups 9 and 12 and ITU-R Study Group 6. This organization is composed of experts from various backgrounds and affiliations, including participants from several internationally recognized organizations working in the field of video quality assessment. Currently, VQEG is conducting an evaluation of metrics in a "multimedia scenario, which is targeted at lower bitrates and smaller frame sizes as well as a wider range of codecs and transmission conditions."

Video quality evaluation has thus become a relevant subject, which is also evidenced by the number of publications, products available (e.g., video quality evaluation probes, known as Witbe robots, for measuring the quality of service offered by multimedia companies such as Portugal Telecom with MEO) and recent international conferences.

Evaluation of video quality can be achieved by two different ways: through subjective and objective metrics. The subjective video quality assessment is recognized as the most reliable mean of quantifying user perception since human beings are the ultimate receivers in most applications. The Mean Opinion Score (MOS), which is a subjective quality measurement obtained from a group of viewers, has been regarded for many years as the most consistent form of quality measurement. However, this quality measurement has some disadvantages. These disadvantages are related with the fact that the MOS method is highly time consuming for most applications and cannot be executed automatically.

In order to provide an automatic evaluation and monitoring of video data quality, reliable and objective metrics are required. By contrast to subjective measurements, the objective quality metrics are based purely on mathematical methods, from quite simplistic ones, like Peak Signal-to-Noise Ratio

[1] http://www.its.bldrdoc.gov/n3/video/standards/index.php

[2] http://www.its.bldrdoc.gov/vqeg/

(PSNR) and the Mean Squared Error (MSE), to sophisticated ones that exploit models of human visual perception and produce results far more consistent with the subjective evaluation [WSB03]. In other words, the objective video quality measurement is done by a software which processes the video signals in order to obtain a video quality score. Thus, this type of video quality metric is more advantageous as it could provide real time quality monitoring for video applications.

Objective video quality metrics can be classified according to the availability of the original video at the quality assessment process. Thus, objective video quality metrics are classified in three classes: Full Reference (FR), Reduced Reference (RR) and No Reference (NR). If the original video is totally available as well as the distorted video, the objective metrics are classified as FR. However, in many practical video service applications the reference video sequences are not accessible; in that case, the metric is classified as NR if it is based only and exclusively on the degraded video. In some cases, to improve the quality estimation, some characteristics of the original video are used, besides the distorted video, thus the objective metrics is categorized as RR metric. Comparatively to FR, few approaches were proposed for RR video quality assessment and even less for NR quality evaluation.

Relatively to FR metrics, it should be highlighted the great effort made by VQEG in order to develop them. VQEG developed FR metrics in two phases. From 1997 to 2000, VQEG carried out the first phase named as “Full Reference Television - Phase I” (FRTV), while the second phase was carried out from 2000 to 2004. Similarly to the first phase, this second phase was named as “Full Reference Television - Phase II” (FRTV) [ITU04]. VQEG begun the RR and the NR television tests in 2000 and it was restarted in 2005. The recently completed Multimedia Phase I test of VQEG assessed the performance of full-reference and reduced-reference perceptual video quality measurement algorithms for QCIF, CIF and VGA formats. Based on it, two standards have been issued by ITU-T [ITU08a] [ITU08b]. As future directions, VQEG propose to study “hybrid” metrics, which look not only at the decoded video as in the other tests, but also at the encoded bitstream[3].

The work presented in this thesis has been organized taking into account the two video quality assessment metrics mentioned previously, the subjective and the objective metrics. The subjective tests have been conducted in order to obtain the MOS of a number of representatives (in terms of spatial features, motion and coding artefacts) compressed video sequences. Two different compression standards, the MPEG-2 and the H.264/AVC, have been considered.

Beyond the fact that subjective evaluation is the most recognized method for quantification of perceived quality, the attainment of the MOS for a number of representative compressed video sequences contributed to build a database of great interest for those working on the video quality evaluation field. The main reason of that significance is due to the fact that the majority of subjective results (e.g. those produced in MPEG groups) are only available for a restrict group of persons. Thus, the production of a database of video sequences and associated MOS, becomes a relevant subject

[3] http://portal.etsi.org/docbox/Workshop/2008/2008_06_STQWORKSHOP/VQEG_ArthurWebster.pdf

since the subjective results as well as all type of information related with them, can be used in future works by people who has interest in video quality evaluation.

The video sequences selected to be presented in the subjective tests session, were compressed using two broadly used compression standards, the MPEG-2 and the H.264/AVC. The MPEG-2 was chosen since it is still widely used as the format of digital television signals. However, H.264/AVC is experiencing a widespread adoption within several countries and covering a wide number of applications ranging from TV broadcast to video for mobile devices and IPTV services.

After the subjective tests having been carried out, a new NR objective quality evaluation method is proposed and evaluated, the main purpose of which is to provide quality scores well correlated with the ones resulting from the subjective tests (MOS).

The present thesis is structured in five chapters.

Chapter 2 provides a general overview of the subjective and objective video quality evaluation metrics.

Chapter 3 presents an overall description of the conditions and choices taken in order to perform the subjective tests sessions, as well as their results. In this chapter, it is also explained how the observer validation should be conducted with the aim of guaranteeing the reliability of the subjective tests results.

In Chapter 4, a new NR objective video quality assessment method is proposed and evaluated.

In Chapter 5, final remarks of the work carried out are presented and future research directions are pointed out.

Chapter 2

Video Quality Evaluation Metrics

2.1 Introduction

As mentioned in the previous chapter, there are, in general, two classes of methods available to measure video quality: the subjective quality metrics and the objective quality metrics. The subjective quality assessment aims to capture, through video's presentations, the user's perception of quality being the most reliable mean of quantifying video quality. It is also the most efficient method to test the performance of human vision models and objective quality assessment metrics. On the other hand, objective quality metrics are based purely on mathematical methods. The goal of this kind of video quality assessment measurements is to design quality metrics that can predict perceived video quality automatically. However, perceived video quality prediction is a difficult task, due to the complexity of the Human Visual System (HVS). This chapter provides a general overview of the two classes of methods mentioned above, namely subjective and objective quality metrics, giving a

particular emphasis to the main characteristics of them.

2.2 Subjective quality metrics

This section presents an overview about methodologies, categories of subjects and rules for performing and designing subjective tests, described and standardized in the Recommendation ITU-R BT.500 and in the Recommendation ITU-T P.910 by the International Telecommunication Union group. The Recommendation ITU-R BT.500 ("Methodology for the subjective assessment of the quality of television pictures") [ITU98] is the reference for anyone who has to deal with quality of video. In this recommendation, different test methods are presented, covering all the possible cases in which visual quality has to be measured.

With regard to the Recommendation ITU-T P.910 ("Subjective video quality assessment methods for multimedia applications") [ITU99], this recommendation describes non-interactive subjective assessment methods for evaluating the one-way overall video for multimedia applications such as videoconferencing, storage and retrieval applications, tele-medical applications, among others. The main difference between these two Recommendations is the fact that the Recommendation ITU-R BT.500 is focused on subjective assessment of video quality for television pictures, *i.e.*, for large video formats; instead, the Recommendation ITU-T P.910 is focused on subjective assessment of video quality for reduced picture formats.

2.2.1 Viewing conditions

Different environments with different viewing conditions can affect the experimental results. Specially, there are three factors that must be considered when performing the subjective tests: the lighting, the ambience noise and the quality and calibration of the display. According to [ITU99], the test should be carried out under the viewing conditions presented in Table 2.1.

Table 2.1: General viewing conditions [ITU99]

Viewing conditions	
Parameters	**Settings**
Viewing distance	1-8H
Background room illumination	≤ 20 lux
Peak luminance of the screen	100-200 cd/m
Ratio of luminance of inactive screen to peak luminance	$\leq 0{,}05$
Ratio of luminance of background behind the display to peak of luminance	$\leq 0{,}2$

In this table, H is the image height.

2.2.2 Selection of test materials

The subjective quality assessment results strongly depend on the videos' scene or sequence content selected to be viewed by the observers. In consequence, the selection of test material must be done carefully.

In order to get meaningful and realistic tests' results, it is important that a wide variety of video material is used during the tests. With regard to test material that should be included in the subjective tests, it is important to incorporate critical material (e.g., videos with more detailed background instead of only homogeneous backgrounds). The main reason for that option is because it is not possible to extrapolate the test results from material that is non-critical, since it is not possible to guess the observers' behaviour under other circumstances. In particular, there are two relevant parameters which should be taken into account when choosing the test scenes: the spatial and the temporal perceptual information of the videos.

In accordance with [ITU99], in order to avoid boring the observers and to achieve a minimum reliability of the results, at least four different types of scenes in terms of spatio-temporal content, should be chosen for the sequences.

2.2.3 Observers selection

The observers' selection is another important task in the subjective quality assessment. In order to produce reliable and coherent results, in accordance with [ITU98], at least 15 observers are needed, with increasing results accuracy and consistency when this number increases. Before performing the subjective tests, the observers should be submitted to ophthalmologic tests in which they are screened for acuity, color blindness and other visual anomalies.

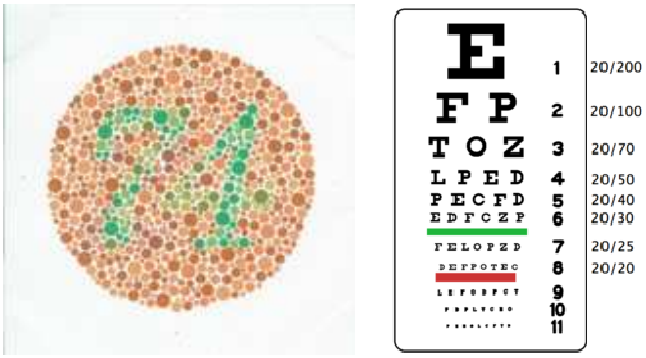
- **Experts or non-experts**

In order to answer to this subject, it should be referred that, in general, the public which consume the video material are commonly non-expert. In short, non-experts make part of the most representative target group comparatively with the experts group. So, in this way, it is obvious that the observers in the subjective quality assessment session should be non-experts. Other reason that supports this choice is directly attached with the fact that the non-experts are not concerned with television picture as part of their normal work. Therefore, the non-experts do not have a pre-determined way of watching a video sequence as the experts have. In [ITU98] preliminary findings suggest that non-experts observers may yield more critical results with exposure to higher quality transmission and displays technologies.

- **Screening the observers**

In the subjective quality assessment, the human eye has a special importance since is through

this mean that the observers will assess the video quality presented in subjective tests session. In this sense, it is necessary to guarantee that the observers are screened in accordance with two main factors before the subjective tests: colour blindness and visual acuity. The colour blindness, or colour vision deficiency, in humans, is the inability to perceive differences between some of the colours that the most common people can distinguish. The visual acuity is the acuteness or clearness of vision, which is dependent on the sharpness of the retinal focus within the eye and is a quantitative measure of the ability to identify black symbols on a white background at a standardized distance as the size of the symbols is varied. The most used standardized tests to evaluate the colour blindness are the Ishiara's test, while the Snellen Eye Chart is used to assess visual acuity (Figure 2.1).



(a) **(b)**

Figure 2.1: (a) Ishiara Test plate; (b) Snellen Eye Chart

The Ishiara's test consists in showing to the observer a set of Ishiara's plates, and in asking him which number he can see inside of each plate. On the other hand, the Snellen Eye Chart is based in the capacity that an observer has to identify a set of letters, at a pre-defined distance from the Chart.

2.2.4 Video evaluation session

According to [ITU98], the quality assessment sessions should not exceed half an hour, since if this does not happen the observer gets tired and, as consequence, the results will not be coherent. These evaluation sessions are divided in two parts: warm-up session and the real test session.

The warm-up session is presented to the observers initially, before the real test session begins, as can be seen in Figure 2.2. The warm-up phase presents the observer with some stabilization presentations. These video sequences are shown with the intention to guarantee the observer's opinion stabilization and to define in his/her mind some video quality boundaries. It is also important to add that the data issued from these presentations should not be taken into consideration for further analysis. After this initial stage has been carried out, the real test session is ready to start and the results from this second phase are the major goal of all entire subjective quality evaluation sessions.

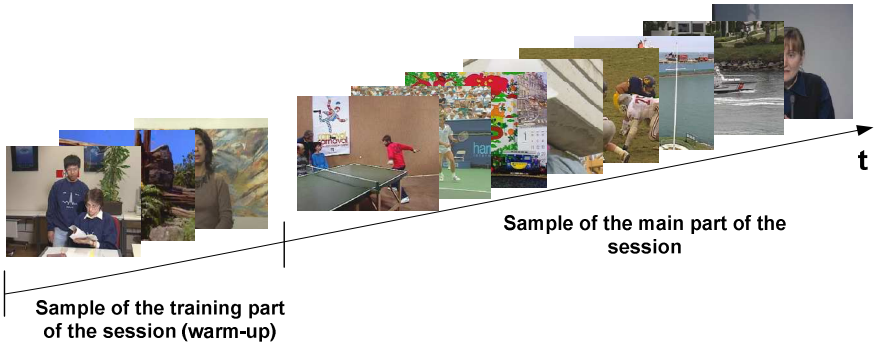


Figure 2.2: Test session structure

It is from this second stage, that the observers' results will be taken into account in order to calculate the Mean Opinion Score (MOS). During this phase it is presented to the observers a set of carefully selected video sequences. The tests can be either single or double presentation: if the reference video and test video are presented only once this presentation is named as single; by contrast, if the reference video and test video are presented twice, the presentation is named as double. This option will be influenced by the test method adopted to perform the subjective tests. With regard to the trial structure, and depending on the type of video quality assessment method used, the reference video can be presented at first place and the test video at second place (which can be degraded or not relatively to the reference video) or, on contrary, the test video can be presented at first place and the reference video at second place. During the presentation, the video sequences should be in a random order. In contrast, the test condition order should be arranged so that any effects on the grading of tiredness or adaptation are balanced out from session to session ([ITU98]). With the purpose of measuring the observer's coherence some manuals recommend to repeat some sequences presentations.

2.2.5 Useful information for the assessment

During the phase that precedes the subjective quality evaluation session, the observers should be carefully introduced to the method of assessment. Questions as "what is the test about?" and "what is the grading scale?", as well as the sequence time, should be well explained in order that the tests results are not influenced by any misunderstanding. Also, as it was already mentioned in the previous section, before starting with the real subjective session, training sequences representing the range and the kind of impairment to be seen during the session should be shown to the observer.

2.2.6 Video quality assessment methods

During the last years a number of subjective testing methodologies were proposed, some of them were standardized in [ITU98] and in [ITU99], namely:

- The Double Stimulus Impairment Scale (DSIS) or Degradation Category Rating (DCR);
- The Comparison Scale Method or Pair Comparison method (PC);

- The Single Stimulus Method (SS) or Absolute Category Rating (ACR);
- The Single Stimulus Continuous Quality Evaluation (SSCQE);
- The Double Stimulus Continuous Quality Scale (DSCQS);
- The Simultaneous Double Stimulus for Continuous Evaluation (SDSCE).

In order to give an overview about the video quality assessment standards depicted above, in the next sub-sections these methods are described giving particular attention to the methodologies and trial structure followed by them.

2.2.6.1 Double Stimulus Impairment Scale (DSIS) or Degradation Category Rating (DCR)

The *Double Stimulus Impairment Scale* (DSIS) [ITU98] is a very useful tool for evaluating clearly visible impairments, such as blockiness, blurring and ringing, which are usually caused by the encoding process. In the context of multimedia applications, this method is equivalent to the *Degradation Category Rating* (DCR) method described in [ITU99]. Furthermore, the DCR method is a key method for the assessment of television pictures whose typical quality represents the highest quality levels found in videotelephony and videoconferencing services [ITU99].

DSIS is not recommended for the quality evaluation of video transmission over packet networks like the Internet. The reason for that, according to Miras et al. [Mira02], is because of packet networks non-deterministic behaviour and the bursty nature of encoded video. This means that, from the user's point of view, perceived quality can vary significantly over time. In this perspective, Pearson et al. [Pear99] discussed several higher-order effects that influence users' quality ratings when assessing video sequences of extended duration. In order to reduce these types of effects on users' quality assessment, what is needed is a method able to dynamically capture user's opinion as the underlying network conditions or visual content complexity change.

In short, the DSIS method should be used when it is important to check the similarity of the test condition with regard to the reference condition; in addition, it should also be used for high quality system evaluation in the context of multimedia communications.

- **Methodology and Trial Structure**

The DSIS method is appropriate for situations where the tests span the full range of impairments responsible for all visible degradation in the image. The observer is presented with video sequences organized in pairs: the first to be displayed is called the *reference* sequence while the second is called the *test* or *impaired* sequence [GGC01]. The reference is the original, undistorted source sequence while the impaired sequence is a distorted version of the reference (for instance, the result of lossy encoding).

As for the number of presentations of each sequence pair during a test session, two variants are possible:

- *variant I*: each pair reference-test is presented a single time, as is shown in Figure 2.3.a). This means that the observer has only one opportunity to view and to analyse the reference and test sequences;
- *variant II*: each pair is presented two times, as is shown in figure Figure 2.3.b). In contrast with variant I, the observer has two chances to watch and to analyse the reference and test sequences, before doing his judgement.

When reduced pictures formats are used in this assessing method, such as CIF, QCIF or SIF[4], it could be useful to display the reference and test conditions simultaneously on the same monitor.

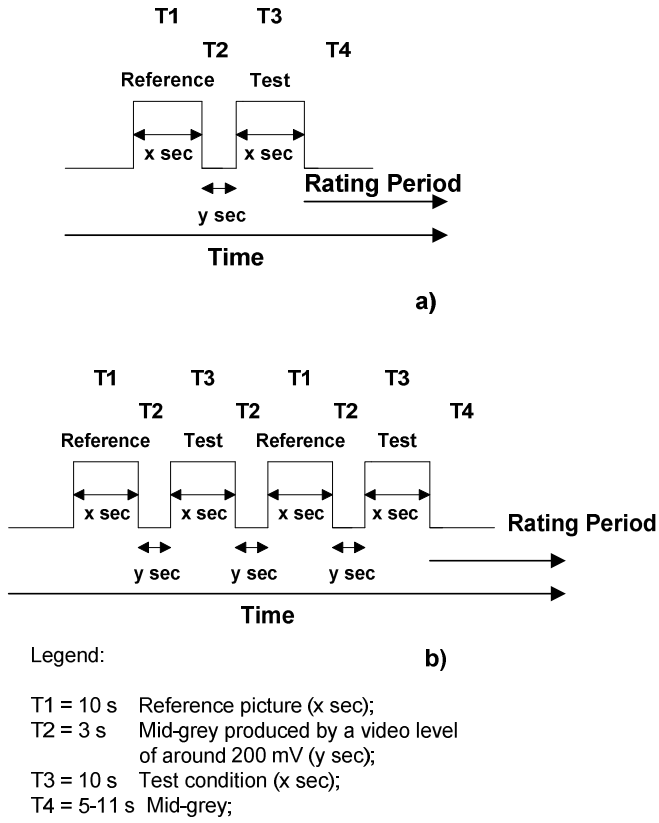


Figure 2.3: Double Stimulus Impairment Test Trial Structure

[ITU98]: a) DSIS I;b) DSIS II

[4] CIF – "*Common Intermediate Format*", typically with a 352 × 288 video spatial resolution.

QCIF – "*Quarter Common Intermediate Format*", typically with a 176 × 144 video spatial resolution.

SIF – "*Source Input Format*", typically with a spatial video resolution of 352 × 240 or 352 × 288.

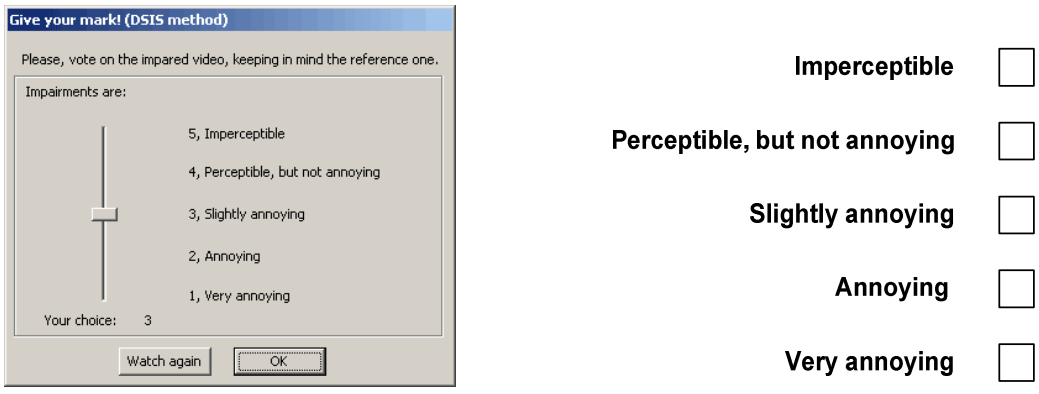
Based on those variants, the DSIS method is known as DSIS I or DSIS II, whenever the method corresponds to the variant I or variant II, respectively. After the sequences have been presented, the observer is asked to vote on the impaired sequence, but keeping in mind the first sequence as reference, in each trial (Figure 2.3).

The DSIS is a method which makes use of five grade impairment scale. The quality assessment grades on this discrete impairment scale are [ITU98]:

- *Imperceptible:* in this case, the test sequence showed to the observer does not seem to be different from the reference sequence;
- *Perceptible, But Not Annoying*: if the observer choose this grade, it is probably because he has noticed some differences between the test and reference sequences, but those differences did not bother him;
- *Slightly Annoying*: the observer sees some degradation in the test sequence, and that degradation bothers him;
- *Annoying*; in this situation, the observer's choice reflects a huge degradation in the test sequence relatively to the reference. This type of degradation bothers so much the observer, that he can stop using this material;
- *Very Annoying*: in this case, the observer is radical on his opinion, *i.e.*, the observer would not watch this kind of material under no circumstances.

It is important to mention that this type of evaluation can be performed using technological means (such as computers) or traditional ways (like paper and pen), as shown in Figure 2.4.(a) and in Figure 2.4.(b), respectively.

The mean opinion scores (MOS) are computed at the end of the session, based in the image quality assessment results given by all observers.



(a) (b)

Figure 2.4: Five point Impairment Rating Scale: (a) using technological means; (b) using traditional ways

2.2.6.2 Comparison Scale Method or Pair Comparison method (PC)

The *Comparison Scale Method* performs a direct head-to-head comparison between two systems (A and B).The purpose of this comparison is to know which system is the best and how much it is better than the other. In accordance with the [ITU99], this method is also addressed to as *Pair Comparison* method in the context of multimedia applications.

- **Methodology and Trial Structure**

The trial structure of this method has the particularity of being blind to the observer, *i.e.*, the reference and test sequences that are shown to the observer are not displayed in a pre-defined order [WR06]. Similarly to other methods, there is the option of presenting each sequence pair once or twice, as depicted in Figure 2.5.(a) and in Figure 2.5.(b).

As for the DSIS method, when reduced resolutions are used (*e.g.* CIF, QCIF or SIF), it could be useful to display each pair of sequences simultaneously on the same monitor [ITU99].

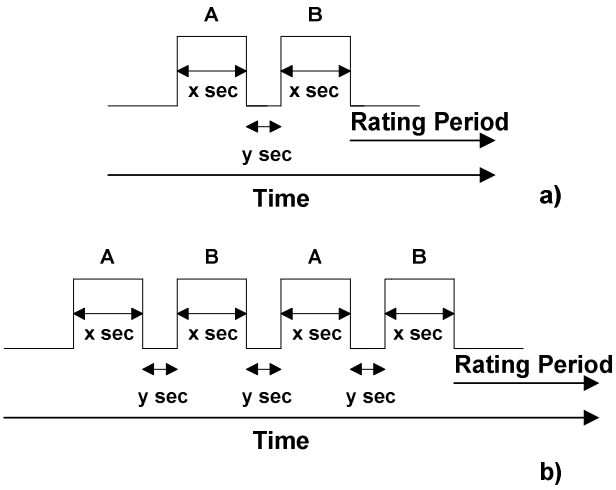


Figure 2.5: Trial structure for Comparison Test for: a) Single presentation; b) Double presentation

With respect to the evaluation scale, in this method the viewers are instructed to assess the difference between the first and second presentations using a 10 cm horizontal scale similar to what is depicted in Figure 2.6. The comparison scale is a continuous scale that has three adjective markers: “A is much better”, “A=B”, “B is much better”.

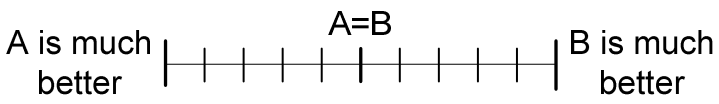


Figure 2.6: Comparison Rating Scale.

2.2.6.3 Single Stimulus Method (SS) or Absolute Category Rating (ACR)

According to [ITU98], the *Single Stimulus* is a method where the test sequences are presented one at a time and are rated independently on a category scale. Using the terminology of [ITU99], the Single Stimulus method is also referred to as *Absolute Category Rating* (ACR). This method allows increasing the observers' time efficiency, since it is fast and easy to implement.

- **Methodology and Trial Structure**

The series of assessment trials should be presented in a random order for each observer. Similarly to the DSIS method and in accordance with [ITU98], it is possible to distinguish two variants based on the presentations' structure, *i.e.,* variant I and variant II.

A typical Single Stimulus trial structure is represented in Figure 2.7.

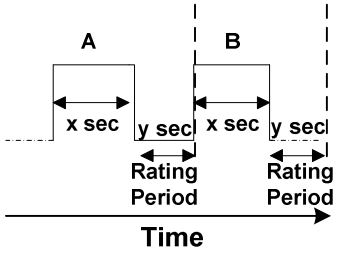


Figure 2.7: Single Stimulus Trial Structure

The subject usually knows the order in which the reference and test video sequences appear in each trial. If the order reference-test is also randomized for each trial, labels "A" and "B" can be used to identify the reference and the test video sequences.

With regard to the evaluation scale, the video quality assessment is performed using one out of four possible scoring scales: a five grade scale, a nine grade scale, an eleven grade scale or a continuous scale with no numbers. The five grade scale, represented in Table 2.2, is the most used one. This numerical scale allows the observer to assign a number to each displayed video sequence that reflects its judgement based on the image quality level.

Table 2.2: Five grade scale

Grading Value	Estimated Quality
5	Excellent
4	Good
3	Fair
2	Poor
1	Bad

However, if more discriminating results are desired, a nine grade scale or even an eleven grade scale should be used. Both are variants of the five grade scale, with additional points for higher discriminative power. Finally, the last scale which can be used in the Single Stimulus method is the continuous scale. This scale enables a non-categorical judgment, for the quality of each image or video sequence. In order to perform his judgment, the observer will mark a point on a line segment that represents the quality scale in which the limits of it represent the worst and the best quality. For reference, the scale usually includes additional quality labels at intermediate points.

2.2.6.4 Single Stimulus Continuous Quality Evaluation (SSCQE)

One of the continuous evaluation methods is the *Single Stimulus Continuous Quality Evaluation* (SSCQE). The SSCQE is a method oriented to the quality assessment in digital television systems. Basically, it consists of measuring the quality of a video sequence along the time, thus the observers are continuously providing their judgment of the video quality on a linear scale. Typically, the assessment material used on this method consists of video sequences that contain scene-dependent and time-varying impairments. In the context of quality monitoring applications, this method yields more representative quality estimates than the previous ones.

- **Methodology and Trial Structure**

The SSCQE methodology belongs to a class of methods where a series of video sequences are presented only once to the observer. This continuous assessment method is the best way to measure the quality variation of a single video clip [Bist05]. In order to take into account the temporal variations of quality, each video sequence should be longer than 10 seconds.

Similarly to other methods that also use continuous rating scales, this method allows the observers to assess both audio and video in video-conferencing applications. In order to keep a high level of concentration and attention from the observers and with the aim to reduce fatigue, the SSCQE method advices the introduction of breaks during each test session. To minimize the contextual effects, the order of the test sequences in the SSCQE is randomized at the clip level, such that every subject will view the test clips in a different order.

With respect to the SSCQE assessment scale, in this method each viewer's opinion is registered twice a second by an electronic handset connected to a computer [Bist05]. The handset is basically a slider mechanism with an associated quality scale, as can be observed in Figure 2.8. Hence, the subject can move the slider to any point over the scale, reflecting his impression of quality at each time instant. The sliding scale is about 10 cm long and is divided in five quality levels.These devices are connected to a computer where the continuous rating of the video material is recorded. It should be noted that each quality label represented in the continuous scale corresponds to a numeric interval. For instance, "Excellent" (100-80), "Good" (79-60), "Fair" (59-40), "Poor" (39-20) and "Bad" (19-0). This association will allow to compute the Mean Opinion Scores (MOS) during the analysis phase.

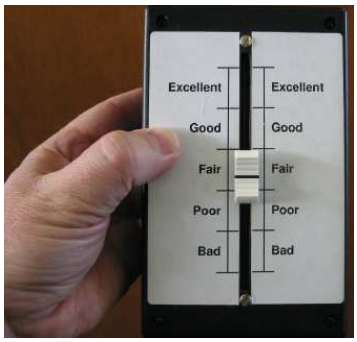


Figure 2.8: Automatic voting device – "Slider" [WP07]

2.2.6.5 Double Stimulus Continuous Quality Scale (DSCQS)

The *Double Stimulus Continuous Quality Scale* (DSCQS) method has been used for performance evaluation of the digital HDTV (High Definition Television) Grand Alliance System, which was the basis for the North American standards for digital TV broadcasting.

This method is especially useful when it is not possible to span the full range of quality stimulus [ITU98]. According to [WP07], the DSCQS method is considered accurate and does not show significant sensitivity to context effects. Context effects occur when subjective scores given by the observer are influenced by the severity and ordering of impairments present in the test material. The DSCQS methodology deals with these by using an alternate way of presenting the video sequences.

➢ Methodology and Trial Structure

In the DSCQS method, the displaying order of the reference and the test sequences is randomized (the reference and test presentations are blind to the subject). Thus, the subject does not know whether the first or the second presentation is the reference or the test sequence. The observer is then asked to evaluate both sequences of images. The DSCQS method can be divided in two variants:

- *Variant I*: Each observer, who is normally alone, is let to switch between the two sequences, A and B, one of which is always the reference and the other is the test.
- *Variant II*: In this variant, it is shown two conditions, A and B, consecutively, to multiple observers, one of which is always the reference and the other is the test.

The reference and test sequences are shown twice to the observer ("double" methodology). After having fully watched both sets of presentations, he is instructed to rate the sequences, as represented in Figure 2.9.

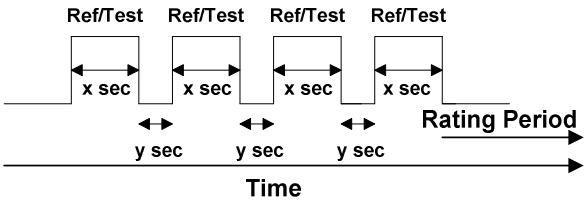


Figure 2.9: Double Stimulus Continuous Quality Scale Trial Structure

In the DSCQS method the viewers are instructed to assess the quality of both presentations using the double vertical scale shown in Figure 2.10. The scale is divided in five equal intervals representing the quality levels. The paper version of the scale should have 10 cm height.

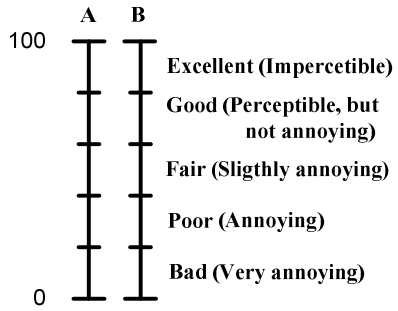


Figure 2.10: Double Stimulus Continuous Quality Scale (parallelism with DSIS's quality adjectives)

After the assessment session, the pairs of quality scores (reference and test) are converted to normalized scores in the range 0 to 100. These scores are spread according to Table 2.3.

Table 2.3: Typical quality assessment scale for DSCQS and SDSCE methods

Perceptual Quality	Equivalent Grade Quality
Excellent	100-80
Good	79-60
Fair	59-40
Poor	39-20
Bad	19-0

After normalization, the differences between the scores given to the reference and to the test sequences are computed for each pair.

It is worth to mention that the use of a continuous scale has the advantage of reducing the amount of quantization error in the observer's responses. In this sense, the DSCQS method is preferred when the quality of the reference and test sequences is similar.

2.2.6.6 The Simultaneous Double Stimulus for Continuous Evaluation (SDSCE)

Besides the other video quality evaluation methods described previously, the *Simultaneous Double Stimulus for Continuous Evaluation* (SDCQE) method is also a standardized and internationally accepted system for image and video quality assessment tests. This video quality assessment method, proposed by MPEG, is suitable to evaluate the effect of sparse impairments, such as transmission errors, on the fidelity of visual information [ITU99]. The SDCQE method, which has been derived from the SSCQE method, differs from that one by making slight deviations in what regards the way of presenting the images to the observers and concerning the rating scale used by them to perform the video quality assessment. According to [ITU98], the SDSCE can be suitably applied to all those cases where fidelity of visual information, affected by time-varying degradation, has to be evaluated.

- **Methodology and Trial Structure**

The SDSCE method consists on assessing two clips simultaneously and continuously. The reference and impaired clips are displayed in parallel positions, as represented in Figure 2.11, using one or two displays (depending on the video resolution). Since the two clips are presented simultaneously, the observer will have to shift his attention between the right and the left presentations, which is a drawback for this method.

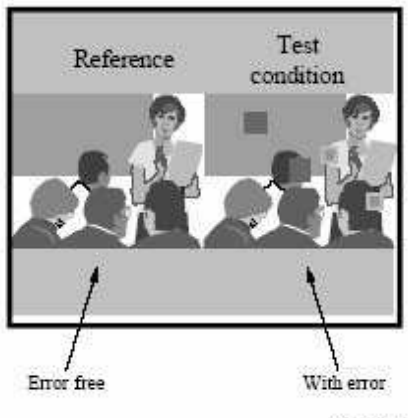


Figure 2.11: SDSCE principle [ITU08]

The observers are asked to continuously judge the fidelity of the video information of the impaired sequence with respect to the reference, by moving a slider on a handset-voting device. The subjects are aware of which sequence is the reference and which sequence is the test condition.

The SDSCE, like other continuous methods previously described, provides to the subjects a continuous scale. The grade given by them will measure indirectly the level of the impairment in the test condition comparatively to the reference condition quality. Once again, the assessment scale associated to this type of method is divided into five equal intervals. These five intervals will correspond to five different qualitative adjectives, presented in Table 2.3.

Similarly to the SSCQE method, the SDSCE method provides to the viewers a device named "slider". It will be through this automatic voting device that the observers will give their perceptual quality opinion.

2.2.6.7 Main Characteristics of Subjective Methods

Based on the description of the different subjective methods, it is possible to summarize, as shown in Table 2.4, the main features of them.

Table 2.4: Video quality assessment methods' main features

Parameter	DSIS / DCR	DSCQS	SSCQE	SDSCE
Selection of test methods	- To measure the robustness of systems (DSIS); - When is testing the fidelity of transmission (distorted video) with respect to the reference signal (DCR);	-To measure the quality of systems relative to a reference; -To measure the quality of a stereoscopic image coding;	- To measure video quality in digital television systems; - The best way to measure the quality variation of a single video clip;	- To measure the fidelity between two impaired video sequences; - To compare different error resilience tools;
Explicit reference	Yes	No	No	Yes
Hidden reference	No	Yes	No	No
Scale	Very annoying to imperceptible	Bad to excellent	Bad to excellent	Bad to excellent
Sequence length	10s	10s	5min	10s
Two simultaneous stimuli	Yes [(1)]	Yes	No	Yes
Presentation of test material	Variant I: once Variant II: twice shown consecutively	Twice shown consecutively	Once	Once
Videos per trial	2	2	1	2
Voting	Only test sequence	Test sequence and reference	Test sequences	Difference between the test sequence and the reference simultaneously shown
Continuous quality evaluation along time	No	No	Yes (moving the slider in a continuous way)	Yes (moving the slider in a continuous way)
Display	All (mainly TV)	All (mainly TV,DLP[(2)])	All (mainly TV)	All (mainly TV)

[(1)] *According to ITU-T P910, it is possible to use a simultaneous presentation when using a reduced picture format, like CIF, QCIF,SIF (for DCR method);* [(2)] *Digital Light Processing which represents a technology used in projectors and video projectors.*

In accordance with the methods' description provided in this chapter, it is not possible to definitively recommend one method over the others, since all have strengths and weaknesses. Therefore, the experimenter who is leading the test session should select the method which he thinks

that it is more adequate for the circumstances.

2.3 Objective quality metrics

This section overviews the main characteristics of objective quality metrics and presents some state-of-the-art metrics to perform this type of video evaluation.

2.3.1 Classification of objective metrics

The objective quality metric should allow to obtain a good prediction of the video quality scores that human observers would give to that video sequence. These video quality metrics can provide quality control of the compressed video and more generally Quality of Service (QoS) in video communications.

They can be categorized in three classes, based on the amount of information about the reference video required and available to estimate the video quality:

- *Full Reference metrics (FR)*: These metrics require the original video and the distorted video;
- *Reduced Reference metrics (RR)*: These metrics require the description of some parameters from the original video and the distorted video;
- *No Reference metrics (NR)*: In contrast to FR and RR, these kinds of metrics only need the distorted video.

The FR metrics are the most studied and developed objective metrics. This type of metrics may have test implementations and, at the same time, may provide good results respectively to the fidelity of video. Typically they are based in a frame-by-frame comparison between the reference and the distorted video sequences, requiring an accurate spatial and also temporal alignment of the two videos, which may be difficult to achieve in practise. This spatial/temporal alignment requirement of the two videos is important since every pixel in every frame of the distorted video must be matched with its counterpart in the reference video, in order to allow a perfect frame-by-frame comparison among them.

In a NR objective metric scenario, the quality scores prediction is obtained through the information available in the receiver side only. Contrarily to FR metrics, in NR metrics there is no need to enforce a spatial and temporal alignment of the reference and distorted videos since no frame-by-frame comparison of both videos is performed.

The major drawback of this kind of metrics, in accordance with [Wink07], is related with the fact that the NR metrics lies in telling distortions apart from content, a distinction humans are typically able to make from experience. In the case of the NR metrics, it is necessary to begin by making assumptions about two important topics, the video content and/or the distortions of interest. As result

of these suppositions, the rise of the risk of confusing actual content with distortions is a reality.

In literature, a limited number of NR metrics has been proposed. However, recently this topic has attracted a great deal of attention. As example of that, is the fact that the VQEG considers the standardization of NR video quality evaluation methods as one of its future working direction. Proposed NR algorithms falls typically in two categories of methods: those that evaluate some specific coding artefacts, such as block effect in block-based DCT compression methods, edge discontinuities, etc; those that estimate pixels distortion and weight those distortions according to some human visual model.

In a RR metric scenario certain features or physical measures are extracted from the original video and then transmitted to the receiver as side information in order to help evaluating the quality of the video. Thus, this class of metrics will require additional bandwidth (or additional channel) to send the side information. Similarly to the FR metrics, the RR metrics may also require a spatial and temporal alignment between the side information and the distorted videos; however, this process is normally less demanding than in the FR metrics, since in this case only the extracted features from the reference video need to be aligned.

The development of RR as well as NR metrics systems has become a priority matter to the video quality community, since in the context of video distribution scenario it is desirable to perform quality evaluation at the receiving side with low-level of information or specially without accessing any information from the original media data.

2.3.2 Objective assessment approaches

According to [Wink07], the measurement of the video distortions in a video communication system can be performed in two ways:

- *Data metrics*: In order to measure the amount of distortion introduced by the capture, compression and transmission processes, these metrics take into account only the signal reliability without considering the content of the video under analysis.
- *Picture metrics*: This distortion measurement is focused on the content of the video under analysis, *i.e.*, this approach allows quantifying the effect of distortions and content on perceived quality. In this case, these metrics are closer to the human perceived quality than the *Data metrics* method.

The most relevant example of a simple data metric is the MSE, or its equivalent PSNR. Although it does not correlate well with the subjective evaluation, it is widely used in video quality evaluation mainly due to its computational simplicity.

A good example of the MSE limitation is the fact that two videos, quantitatively with the same MSE values, can in fact have different subjective scores. The MSE does not make any distinction on the different types of artefacts, *i.e.*, MSE treats all errors in the same way, regardless of its influence on the video's quality.

However, the MSE/PSNR has a good performance when comparing two compressed versions of the same original video sequence, using the same encoder. This phenomenon occurs because the compressed videos are being encoded with the same distortion characteristics. Other advantage of using the MSE/PSNR, according to Brandão et al. [BQ08b], is the fact that this video quality metric can be used as a NR quality metric, *i.e.*, it is possible to produce accurate PSNR estimates without the need of the original data.

Formally, the MSE is given by

$$\mathrm{MSE} = \frac{\sum_{i=1}^{M}\sum_{j=1}^{N}\left[f(i,j) - F(i,j)\right]^2}{M \times N} \qquad (2.1)$$

where,

$f(i,j)$ is the original video component (luminance or chrominance) at pixel (i,j);

$F(i,j)$ is the distorted video component at pixel (i,j);

M is the picture width;

N is the picture height.

The PSNR is derived by setting the MSE in relation to the maximum possible value of luminance (for a typical 8-bit value this is $2^8 - 1 = 255$), and is usually expressed in logarithmic units through:

$$\mathrm{PSNR} = 10\log_{10}\left(\frac{255^2}{\mathrm{MSE}}\right), \text{ [dB]} \qquad (2.2)$$

Picture metrics are the result of much effort made in order to develop better visual quality metrics that quantify the effects of distortions and content on perceived quality. As a consequence of this effort, the *Picture metrics* can be classified in two groups [Wink07]:

- a vision modelling approach;
- an engineering approach.

The vision modelling approach is based particularly on HVS, *i.e.*, these kinds of metrics try to include human vision characteristics which seem to be relevant to picture quality, like contrast sensitivity, color perception, applying models and data from psychophysical experiments.

The engineering approach, instead of focusing in the HVS as it is done in the vision modelling approach, relies on the extraction of certain features or artifacts in the video under analysis. This type of approach, which has gained popularity in recent years, focuses on the strength of these extracted

features and then takes them into account in order to estimate the overall quality of it. The extracted features are image structural elements or specific distortions, introduced by a particular video processing step, compression technology, or transmission link.

An example of the engineering approach, which has gained high popularity in recent years, is the Structural Similarity Index Method (SSIM) [WBSS04]. In this metric, video degradations are considered as perceived structural information loss instead of perceived errors.

The SSIM consists in computing from the original and from the distorted video, three measurements: luminance, contrast and structural distortions. These measurements are then separately compared. At the end, the comparison results are combined to yield an overall similarity measure. Figure 2.12 describes briefly the SSIM's process.

The SSIM's measurement system has been constructed on the assumption that video degradation is often caused by the loss of underlying structured information. One of the strengths of this video quality assessment metric is the fact that this metric has showed to perform well for some artifacts which are not directly related to the compression step, such as added noise. On the other hand, the SSIM can also adapt to artifacts which are directly related to low bitrate video compression (such as blocking effect) and provide as well perceptually consistent quality predictions.

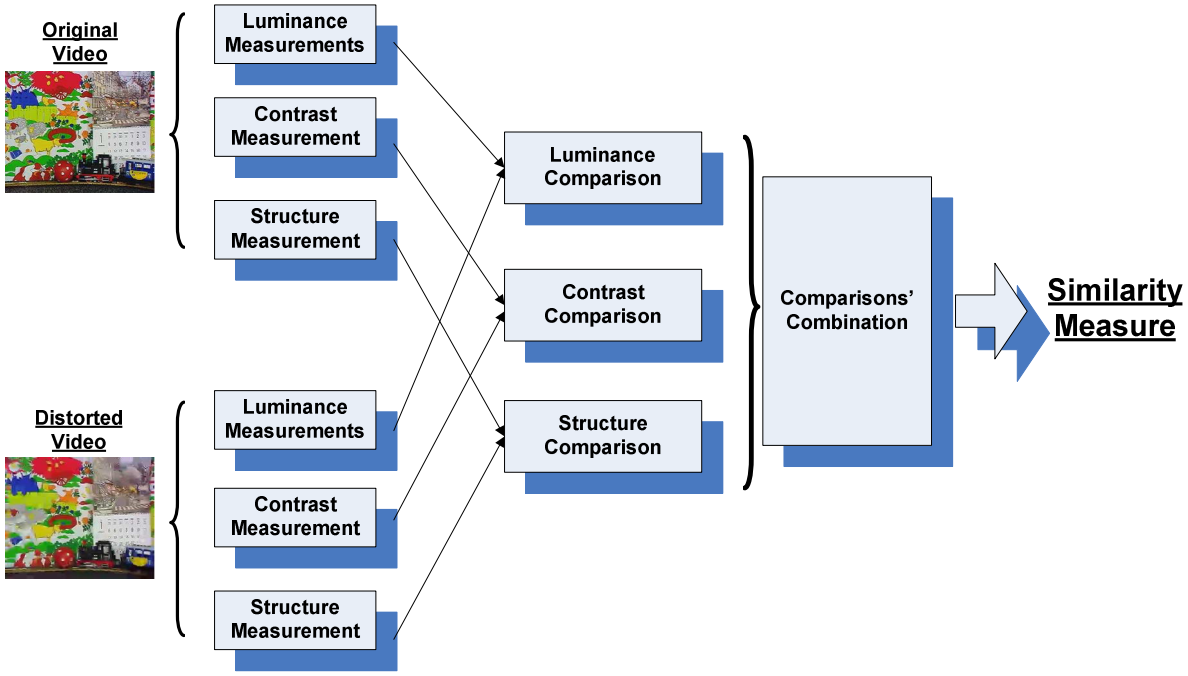


Figure 2.12: SSIM's measurement system

Although the SSIM presents a relative simplicity, this metric behaved quite well on the VQEG FR-TV Phase I database.

Besides the SSIM method, there are other popular structural information based metrics, such as the metric developed by Wolf and Pinson, named as Video Quality Metric (VQM). The VQM metric, similarly to the SSIM extracts, from the distorted video, a restrict set of features, which are selected empirically and carefully from a group of possible features. After that selection phase, those features

are then compared analogously with the features from the reference video. According to [Wink07], the VQM was among the best metrics in the VQEG FR-TV Phase II evaluation. Another example of this type of approach is the metric designed by Hekstra et al. [HBL02], named Perceptual Video Quality Measure (PVQM). This video quality assessment uses a linear combination of three particular features, which are the loss of edge sharpness, the color error normalized by the saturation, as well as the temporal variability from the reference video. In accordance with [Wink07], the PVQM was also one of the best metrics in the VQEG FR-TV Phase I test. All the above mentioned metrics – SSIM, VQM and PVQM – belong to the class of FR metrics.

Rec. ITU-T J.247 [ITU08b] provides four FR video quality estimation methods:

- NTT Full Reference Method (developed in Japan);
- OPTICOM's Video Quality Method (developed in Germany);
- Psytechnics Full Reference Method (developed in United Kingdom);
- Yonsei University Full Reference Method (developed in Korea).

The NTT full reference prediction model, estimates subjective video quality by an alignment process and a video quality algorithm that reflects human visual characteristics. The second video quality model, developed in Germany, also known as Perceptual Evaluation of Video Quality (PEVQ) model, is a very robust one which was designed to predict the effects of transmission impairments on the video quality as perceived by a human subject. The main targets of this model are mobile applications as well as multimedia applications. Regarding the Psytechnics full-reference video quality assessment algorithm, it consists on the identification of perceptually relevant boundaries and in the inclusion of a model of the human visual system. These two elements allow the model to identify and quantify errors perceived by human viewers. As a result, the Psytechnics video quality assessment model produces (objective) quality predictions that correlate highly with human (subjective) quality judgment.

It is observed that the human visual system is sensitive to degradation around the edges. Furthermore it is observed that video compression algorithms tend to produce more artefacts around edge areas than on the remaining ones. Based on this observation, the Yonsei University full reference model provides an objective video quality measurement method that measures degradation around the edges. In this model, an edge detection algorithm is first applied to the source video sequence to locate the edge areas. Then, the degradation of those edge areas is measured by computing the mean squared error. From this mean squared error, the Edge PSNR (EPSNR) is computed. Furthermore, the model computes two additional features which are combined with the EPSNR to produce the final video quality metric.

The full reference method is generally accepted as the model that provides the best accuracy for perceptual picture quality measurements. However it is also known that this method is only suitable when the original (reference) and the distorted video are totally available at the receiver side, and consequently this type of architecture are not adequate for practical media distribution scenarios.

In the context of multimedia distribution scenarios, it is desirable to track media quality at the receivers. This could enable new services, such as users paying proportionally to the quality they get, and new server possibilities, such as adjustment of streaming parameters as a function of the perceived quality. Thus, in order to assess video quality without requiring the original video data, Reduced Reference and No Reference methods (NR) are required.

As it was mentioned in sub-section 2.3.1, the RR measurement method can be used when features extracted from the unimpaired reference video signal are readily available at the receiver side. Based on that, Rec. ITU-T J.246 [ITU08a] proposes some RR models, based on the measurement of the edges degradation. According to these models, an edge detection algorithm is first applied to the source video sequence to locate the edge pixels and then, the degradation of those edge pixels is measured by computing the mean squared error (2.1), i.e., the amount by which the edge pixels from the source video sequence differs from the ones located on the distorted video is quantized. After, the edge PSNR is computed. Depending on the nature of videos and compression algorithms, a different edge detection algorithm can be chosen [ITU08a].

In [OD07], another RR video quality metric for AVC/H.264 was proposed. In this case, the RR model evaluates a set of features such as blur or blocking and combines these measurements with few additional data (extracted from the original video and transmitted as side information) into one quality score using multivariate data analysis.

In what concerns the NR metrics, few approaches were proposed in literature and none has been standardized yet. In fact, as it was mentioned before, as future work VQEG considers the standardization of NR video quality evaluation methods as a priority matter. Recently, Brandão [BQ08a] presented an approach that can be used for image quality evaluation without requiring any knowledge about the original signal, thus belonging to the NR image quality metrics category. Quality scores rely on statistical properties of the original, block-based, DCT (discrete cosine transform) coefficient data that are estimated from the received (and quantized) DCT coefficients, and on the perceptual characteristics of the human eye. The main goal was to estimate distortion errors and corresponding perceptual weights, in such a way that quality scores given to the distorted images resemble the perceptual metric proposed by Watson in [Wats93]. The method proposed in [BQ08a] for JPEG encoded images was partially extended to H.264/AVC encoded video in [BQ08b].

Chapter 3

Subjective Quality Evaluation

3.1 Introduction

The subjective quality assessment is a human perception based method that uses structured experimental designs as well as human participants. The goal of these participants is to assess the video quality presented during the subjective quality evaluation sessions. In this chapter, the Mean Opinion Score (MOS) of a number of video sequences is obtained. The MOS is initially computed taking into consideration all the observers present in the subjective tests. In order to guarantee the coherence and the consistency of the results provided by the subjective tests, a statistical analysis is followed with the aim of validate the observers' opinions. After the observer's validation has been

performed the final MOS values are computed. By contrast with MOS initially obtained, the new MOS is calculated taking into account only the coherent observers. At the end of this chapter, the test results are presented as well as the graphics which summarize these results.

3.2 Subjective assessment

This section presents the test methodology used to conduct the subjective tests and the main reasons to follow it. Taking into account sub-section 2.2.6, where subjective video quality evaluation methods were described, as well as [ITU99], in this work the followed method was the Degradation Category Rating (DCR), also known as Double Stimulus Impairment Scale (DSIS). The main reason to choose the DCR was the fact that it is recommended to assess reduced video formats, such as CIF, QCIF or SIF. Furthermore, these reduced video formats are suitable for video applications in 3G wireless networks and for video streaming characterized by low resolutions, and low bitrates; for instance, the CIF and SIF resolutions are commonly used for data-cards and palmtops (PDA), while the QCIF is generally used for cell phones.

3.3 Viewing and test conditions

As mentioned in chapter 2, there are two essential elements for conducting the subjective quality evaluation sessions properly: the environmental viewing conditions and the test conditions. The main test conditions are:

- Maximum test duration per session: 22 minutes
- Maximum number of observers per session: 2
- Total number of observers in the subjective tests session: 22
- Viewing distance: 8 x of the picture height shown in the screen (H)

In Figure 3.1 the testing room used in this dissertation is schematically presented where the parameter H indicates the height of the video shown on the screen.

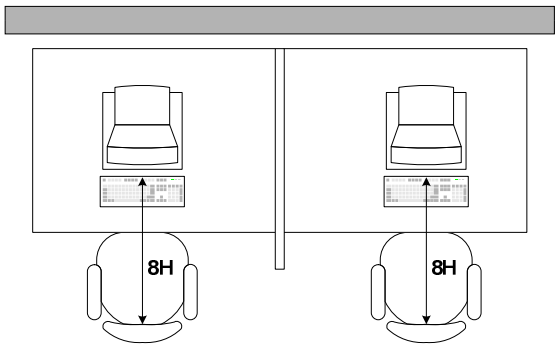


Figure 3.1: Testing room

Table 3.1 presents some aspects related to the display and room characteristics used in this dissertation.

Table 3.1: Display and Room's conditions

Display and Room's conditions	
Parameters	**Settings**
Height of the picture shown in the screen (H)	8 cm
Viewing distance	64 cm
Background room illumination	13,45 lux
Peak luminance of the screen	95,8 lux
Luminance of inactive screen	2,23 lux
Luminance of background behind the display	10,15 lux
Ratio of luminance of inactive screen to peak luminance	0,023
Ratio of luminance of background behind the display to peak of luminance	0,14

Based on Table 3.1, it is possible to conclude that the values achieved for our display and room's conditions are within the values recommended in [ITU99] (see Table 2.1).

3.4 Characterization of the test sequences

When selecting the video sequences to be used in the tests, it is important to take into account the factors that most influence the HVS. According to [RNR07] the human visual perception of video content is determined by the video spatial information, as well as by the type, direction and speed of movement, or temporal activity.

Since a small number of test sequences will be used in the test sessions, it is important to choose a set of sequences that span a large range of possible spatial and temporal information. In other words, the chosen sequences should be well representative of the video sequences that can be encountered in the envisaged application. Hence, in order to choose a set of video sequences, the spatial and temporal activities of each video sequence has to be computed. The literature provides several different methods of measuring these activities. In this work, the methods recommended in [ITU99] have been used.

- **Spatial activity**: The spatial activity measurement uses two filters that work independently of each other. One filter is responsible to compute horizontal pixel differences, or horizontal picture gradient, as shown in Figure 3.2.(a), while the other computes vertical pixels differences, or vertical picture gradient, as shown in Figure 3.2.(b). These two filters are called Sobel filters. Mathematically speaking, the Sobel filtering consists in convolving the two 3x3 kernels presented in Figure 3.2 with each frame of the video sequence. In order to obtain for each pixel a single measure, the gradient norm (the square root of the sum of the vertical and horizontal gradient squares) is computed. Then, the standard deviation of it is obtained in a frame basis. This process is repeated for each frame of the video sequence and results in a time series of spatial information of the scene. In order to achieve a global value for spatial activity, the maximum value in the time series is selected with the purpose of representing the spatial information content of the scene.

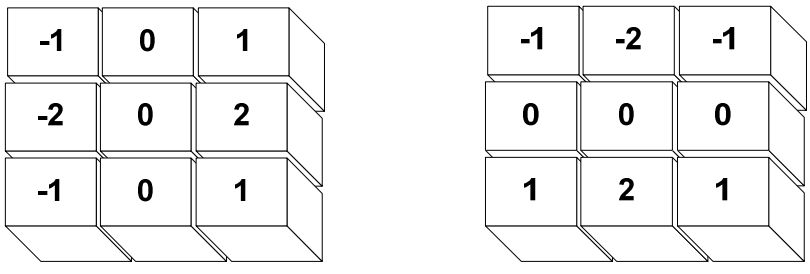


Figure 3.2: Sobel filters. (a) Sobel filter responsible for detecting horizontal pixel differences; (b) Sobel filter responsible for detecting vertical pixel differences [WP99]

Figure 3.3 shows the resulting gradient norm for two frames of the video sequences “Stefan” and “Football”.

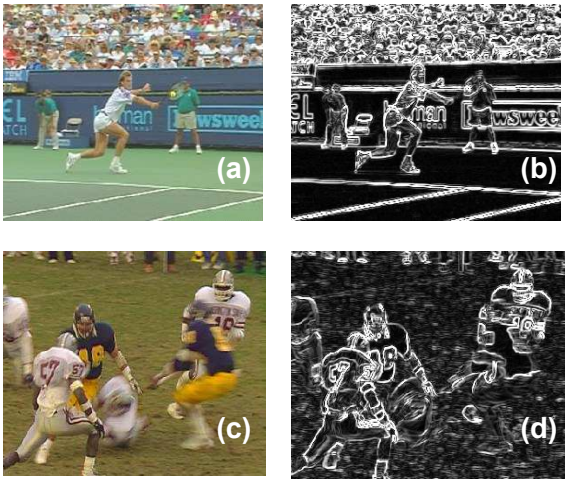


Figure 3.3: (a) and (c) Original video frames; (b) and (d) Corresponding gradient norm images

In Figures 3.3.(b), and (d), higher level of luminance values correspond to higher values of gradient norm (spatial activity).

- **Temporal activity**: According to [ITU99], a temporal activity measure can be obtained computing the difference, pixel by pixel, between each two successive frames of the video sequence. This process is repeated for all video frames. After this procedure has been carried out, the standard deviation of the frames differences is computed. Similarly to what happens in the spatial activity, the global temporal activity value is computed as the maximum of these standard deviations.

Figure 3.4 presents, in the right side, two consecutive frames of the original video and, in the left side, the resulting difference between the two original frames.

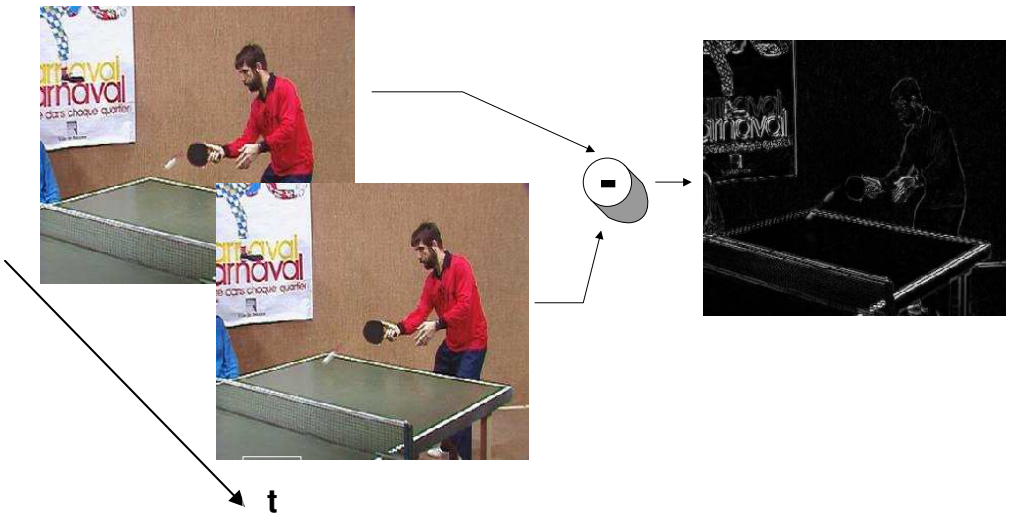


Figure 3.4: Temporal activity measurement process in a video sequence

According to Figure 3.4, two successive frames are first of all compared pixel by pixel, in order to measure the absolute difference between them. After this comparison process has been carried out a luminance frame is created which represents the temporal activity existing between the two compared frames; the higher the temporal activity variation between the two compared frames, the higher will be the luminance content of the frame difference.

3.5 Video sequences selection

According to what was described in the section 3.2, the spatial-temporal information was computed for a set of video sequences, commonly used by the video coding community, in CIF format. The results are presented in Figure 3.5.

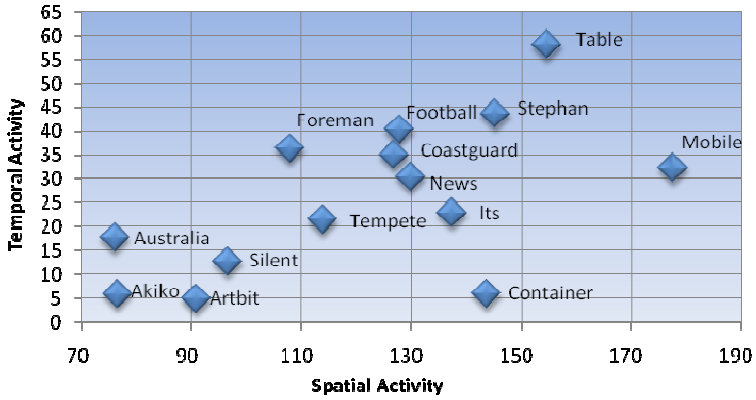
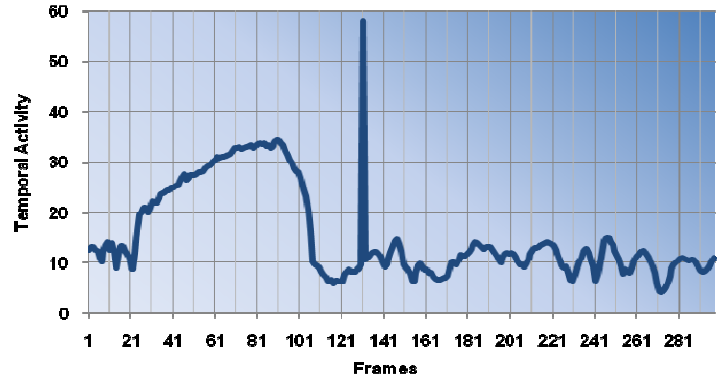
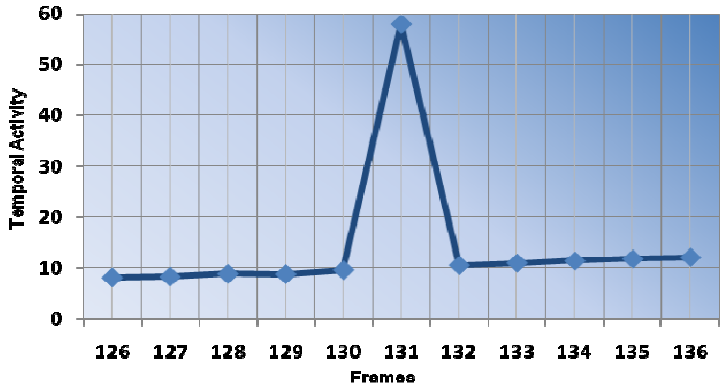


Figure 3.5: Spatial-temporal activity of a video sequence set (CIF format)

However, there is one aspect that should be taken into account. This aspect is related to the fact that some videos present abrupt changes of camera perspective during video acquisition which will consequently cause an abrupt change of scenario in two consecutive frames. Thus, when measuring the temporal activity of a video sequence, the resulting global value may not reveal the true value of the temporal activity. Figure 3.6 presents the temporal activity of the video sequence “Table”, frame by frame.



(a)



(b)

Figure 3.6: “Table” temporal activity, frame by frame considering (a) all the video sequences; (b) the frames affected by abrupt change of camera perspective

In Figure 3.6.a), it is possible to observe that the sequence “Table” presents the symptoms described previously, relatively to temporal activity global value. Analysing the temporal activity evolution of the “Table” sequence, it is possible to see that this video sequence shows, in a great part of the time, a regular temporal activity. However it is also possible to see from Figure 3.6.b) that between the frame number 130 and the frame number 132 there is a sudden peak in temporal activity value. Thus, when measuring the global temporal information of that sequence, there is a discrepancy between the real value of the temporal activity and the computed one. In order to minimize and smooth this effect, it was applied a mathematical procedure, named as percentile 95%, to the temporal and spatial activities of each one of the video sequences selected for the test sessions. Figure 3.7 presents the global results of spatial and temporal activities, after applying the percentile 95% to each video sequence.

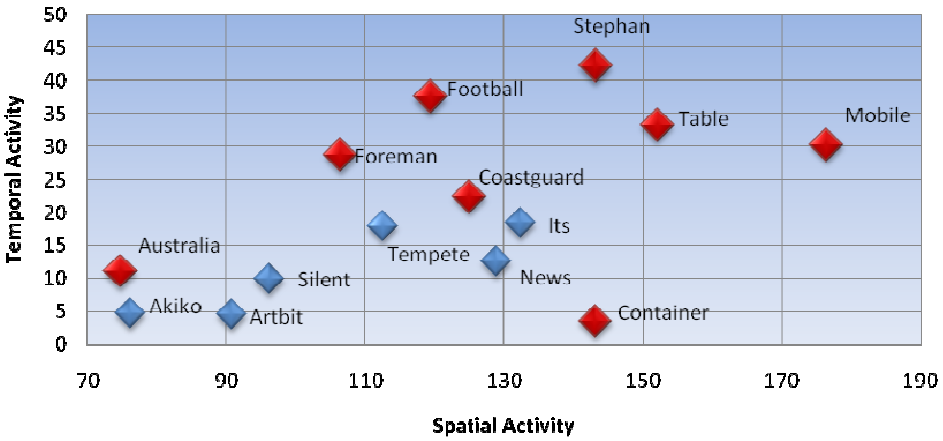


Figure 3.7: Selected video sequences for spatial-temporal activity has been taking with percentile 95%

In accordance with Figure 3.8 and taking into account that the video sequences must span a large portion of the spatial-temporal information, eight video sequences were chosen. Figure 3.8 presents a sample image of each video sequence selected for the test sessions. Concerning the video sequences "Coastguard" and "Football", although presenting similar spatial-temporal activities, in terms of subjective evaluation they can achieve different scores, for the same compression bit-rate. This can be justified by the fact that these video sequences present distinct types of content. So, in order to collect the MOS resulting from different scenarios, both video sequences were selected for the test sessions.

Legend:

1 – "Stephan"
2 – "Coastguard"
3 – "Container"
4 – "Foreman"
5 – "Mobile"
6 – "Table"
7 – "Football"
8 – "Australia"

Figure 3.8: Video's sequences used in the subjective tests

3.6 Video compression

The original video sequences selected in section 3.3 were encoded using two standard video compression techniques: H.264/AVC[5]and MPEG-2[6]. Figure 3.9 shows different levels of video quality degradation that can be found during the subjective quality evaluation tests, and resulting from the compression process. In both standards each sequence was encoded with 4 different bitrates. The reasons for that were to test the HVS perception to different kinds of video qualities and to force the observers to use all rating scale. Therefore, at the end of the session, the MOS would be more consistent and reliable.

Figure 3.9: Video sequences encoded with different values of bitrate

In what concerns the number of test presentations showed to each observer in the subjective tests, and in accordance with section 3.5, they were:

- number of test sequences: 8;
- number of test conditions: 4 different compression bitrates for each video sequence;
- number of test presentations: 32.

[5] H.264/AVC – This video compression standard was developed by the ITU-T Video Coding Experts Group (VCEC) together with the ISO/IEC Moving Picture Experts Group (MPEG), and it was the product of a partnership effort known as the Joint Video Team (JVT).

[6] MPEG-2 – MPEG-2 is a standard for video compression which was developed by the Moving Pictures Expert Group (MPEG).

In the first subjective quality assessment session, the video sequences were encoded using the H.264/AVC compression standard, with the compression bitrates shown in Table 3.2. In the second session, video sequences were encoded using the compression standard MPEG-2, with the compression bitrates shown in Table 3.3. The compression bitrates presented in Table 3.2 and in Table 3.3, were selected with the goal of displaying to the observers different types of video quality for each video sequence. All video sequences used during the tests session had a 352 x 288 spatial resolution, 10 s of time duration and, except for "Australia" (which has a frame rate of 25 Hz), a frame rate of 30 Hz.

According to Table 3.2 and Table 3.3, it is possible to observe that, in a general, the compression rates in H.264/AVC are larger than in MPEG-2 standard. In fact, to achieve similar video quality degradation in H.264 as in MPEG-2, it is necessary to decrease the bitrate in H.264 relatively to the bitrate in MPEG-2. With regard to the compression methods, there are several sub-processes that take place during compression, hence different artifacts are introduced into the media by H.264/AVC and MPEG-2. These compression techniques take advantage of the HVS's characteristics in the sense that they eliminate, from the video, certain data that a common human observer is not sensible to.

Table 3.2: Compression bitrates used in H.264/AVC

H.264	Trial 1	Trial 2	Trial 3	Trial 4
Stephan	128 kbit/s	256 kbit/s	512 kbit/s	1024 kbit/s
Table	64 kbit/s	128 kbit/s	256 kbit/s	512 kbit/s
Mobile	64 kbit/s	128 kbit/s	256 kbit/s	512 kbit/s
Football	256 kbit/s	512 kbit/s	1024 kbit/s	2048 kbit/s
Foreman	64 kbit/s	128 kbit/s	256 kbit/s	512 kbit/s
Coastguard	64 kbit/s	128 kbit/s	256 kbit/s	512 kbit/s
Container	64 kbit/s	128 kbit/s	256 kbit/s	512 kbit/s
Australia	32 kbit/s	64 kbit/s	128 kbit/s	256 kbit/s

Table 3.3: Compression bitrates used in MPEG-2

MPEG-2	Trial 1	Trial 2	Trial 3	Trial 4
Stephan	512 kbit/s	1024 kbit/s	2048 kbit/s	4096 kbit/s
Table	256 kbit/s	512 kbit/s	2048 kbit/s	1024 kbit/s
Mobile	256 kbit/s	512 kbit/s	1024 kbit/s	4096 kbit/s
Football	512 kbit/s	1024 kbit/s	2048 kbit/s	4096 kbit/s
Foreman	256 kbit/s	512 kbit/s	1024 kbit/s	2048 kbit/s
Coastguard	256 kbit/s	512 kbit/s	1024 kbit/s	2048 kbit/s
Container	128 kbit/s	256 kbit/s	512 kbit/s	1024 kbit/s
Australia	128 kbit/s	256 kbit/s	512 kbit/s	1024 kbit/s

However, with increasing levels of compression rates, the distortion introduced by the process may overtake the perceptivity threshold, and consequently introduce visible artifacts into the video. When a video sequence is encoded with H.264, the artifact introduced by this compression is essentially the blur effect (Figure 3.10.a)). Blurriness is caused by the removal or attenuation of high-frequency content due to quantization or low-pass filtering and is characterized mainly by smudging of edges and loss of detail throughout the image. In contrast, when the video sequence is encoded with MPEG-2, the most visible artifact is the block effect (Figure 3.10.b)). In this case, blockiness is characterized by introducing several and visible small blocks conducting to an accentuated image distortion.

(a) **(b)**

Figure 3.10: Artifacts introduced by (a) H.264 compression (blur effect) and (b) MPEG-2 compression (block effect)

3.7 Video quality evaluation program interface

The program used to carry out the subjective tests was the MSU (Moscow State University) perceptual video quality player which was developed by Graphics&Media Lab Video Group. The program interface is shown in Figure 3.11. In order to begin the video quality evaluation session, the observer has to press the start button (d), and then the video sequence, which can be the reference video or the distorted video (b), is displayed on the screen.

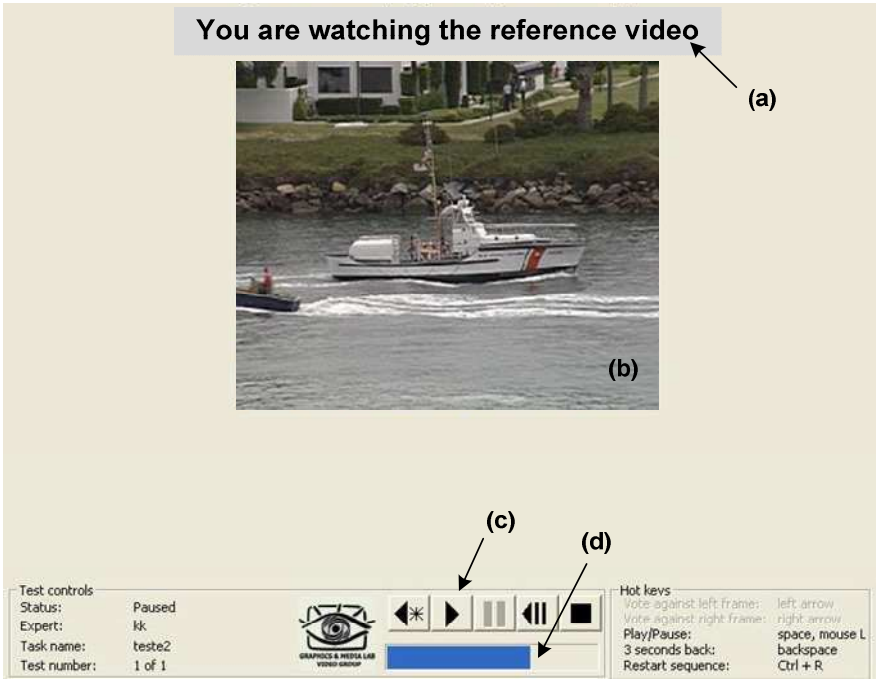


Figure 3.11: MSU perceptual video quality player interface: a) video label (reference/distorted video); b) video window; c) play button to start the video sequence; d) video time bar

Additionally, this interface allows the observer to have a perception of the video time duration through the video time bar (e). After displaying the reference video and the distorted video respectively, a five point impairment rating scale window (Figure 2.4.a)) is shown on the screen where each observer can give his video quality opinion.

3.8 Statistical analysis

After the subjective quality tests have been concluded, the video quality assessment method selected to perform those tests (the DCR/DSIS method), and described in Chapter 2, produced distributions of integer values between 1 and 5. The difference between the observers' opinions about the video quality will result in variations on the observers' scores. This phenomenon is vulgar when working with a group of humans, since the differences in judgement between them, is a constant. With reference to the statistical analysis of the results, the main steps followed were ([ITU98]):

- Calculation of mean scores;
- Calculation of confidence interval;
- Observers validation;
- Calculation of MOS final values.

3.8.1 Calculation of mean scores

The first step of the analysis of the results is the calculation of the mean opinion score, $\overline{u}_{jkr}$ (or MOS), for each of the presentations, which is given by:

$$\overline{u}_{jkr} = \frac{1}{N}\sum_{i=1}^{N} u_{ijkr} \qquad (3.1)$$

where,

u_{ijkr} is the score given by observer i, for test condition j, sequence k, repetition r;

N is the number of observers present in the assessment sessions.

3.8.2 Confidence interval

In order to verify if the distribution of scores from the test presentation is normal or not, the β_2 test was applied. The β_2 test consists in the calculation of the kurtosis coefficient of a function. This kind of test allows to know if the distribution is symmetric or not. Therefore, if the distribution guarantees the β_2 test conditions, it is possible to approximate that distribution by a normal distribution. The $\beta_{2\,jkr}$ coefficient, related with a test condition j, sequence k and repetition r, is given by:

$$\beta_{2\,jkr} = \frac{m_4}{(m_2)^2}$$

where,
$$m_x = \frac{\sum_{i=1}^{N}\left(u_{ijkr} - \overline{u}_{ijkr}\right)^x}{N} \qquad (3.2)$$

If β_2 is between two and four, the distribution can be considered, according to [ITU98], as a normal distribution (also known as a Gaussian distribution).

The mean scores computed for each one of the presentations should have always associated a confidence interval, since it is based on this interval that the reliability of the test results can be

guaranteed. The confidence interval associated to each mean score, is derived from the standard deviation and size of each sample. According to [ITU98], the confidence interval that should be used in this type of analysis is a confidence interval of 95.5%, typically.

The 95.5% confidence interval is given by:

$$[\overline{u}_{jkr} - \delta_{jkr}, \overline{u}_{jkr} + \delta_{jkr}] \quad (3.3)$$

where,

$$\delta_{jkr} = 2\sigma_{jkr} \quad (3.4)$$

and σ_{jkr} is the standard deviation for each presentation, given by

$$\sigma_{jkr} = \sqrt{\sum_{i=1}^{N} \frac{(\overline{u}_{jkr} - u_{ijkr})^2}{N-1}} \quad (3.5)$$

Figure 3.12 presents a normal distribution where a 95.5% confidence interval is signalised.

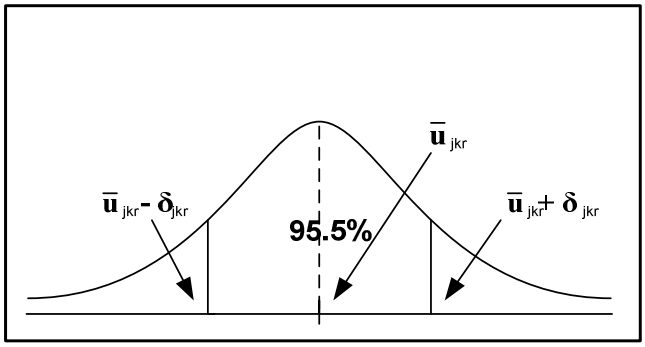


Figure 3.12: Normal distribution interval

According to Figure 3.12, the probability that a random variable X assumes a value in the interval $\overline{u}_{jkr} - \delta_{jkr} < X \leq \overline{u}_{jkr} + \delta_{jkr}$ is equal to 95.5%, *i.e.*,

$$P(\overline{u}_{jkr} - 2\sigma_{jkr} < X \leq \overline{u}_{jkr} + 2\sigma_{jkr}) = 95.5\% \quad (3.6)$$

3.8.3 Observer validation

After having carried out the β_2 test, the scores u_{ijkr} given by each viewer will be compared with the associated outlier condition[7], which can be different if the distribution is normal or not. If the

[7] Outlier - In statistic, an outlier is an observation that is numerically distant from the rest of the data, *i.e.*, an outlier is all observations which are outside of the confidence interval.

distribution of scores follows a normal distribution ($2 \le \beta_{2\,jkr} \le 4$) and the confidence interval is 95.5%, then an observation u_{ijkr} is considered as an outlier if,

$$u_{ijkr} \ge \overline{u}_{jkr} + 2\sigma_{jkr} \quad \text{or} \quad u_{ijkr} \le \overline{u}_{jkr} - 2\sigma_{jkr} \qquad (3.8)$$

On the other hand, if the distribution of scores is not normal, the observation u_{ijkr} is considered as an outlier if,

$$u_{ijkr} \ge \overline{u}_{jkr} + \sqrt{20}\sigma_{jkr} \quad \text{or} \quad u_{ijkr} \le \overline{u}_{jkr} - \sqrt{20}\sigma_{jkr} \qquad (3.9)$$

When the score u_{ijkr} of a viewer is superior to $\overline{u}_{jkr} + 2\sigma_{jkr}$ (if normal) or $\overline{u}_{jkr} + \sqrt{20}\sigma_{jkr}$ (if non-normal), a counter related with that viewer, P_i, will be incremented, *i.e.*,

in case of a normal distribution and $u_{ijkr} \ge \overline{u}_{jkr} + 2\sigma_{jkr}$

in case of a non-normal distribution and $u_{ijkr} \ge \overline{u}_{jkr} + \sqrt{20}\sigma_{jkr}$

$$\Big\} \quad P_i = P_i + 1 \qquad (3.10)$$

On the other hand, if the score u_{ijkr} of a viewer is inferior to $\overline{u}_{jkr} - 2\sigma_{jkr}$ (if normal) or $\overline{u}_{jkr} - \sqrt{20}\sigma_{jkr}$ (if non-normal), a counter associated with that observer, Q_i, will be incremented, *i.e.*,

in case of a normal distribution and $u_{ijkr} \le \overline{u}_{jkr} - 2\sigma_{jkr}$

in case of a non-normal distribution and $u_{ijkr} \le \overline{u}_{jkr} - \sqrt{20}\sigma_{jkr}$

$$\Big\} \quad Q_i = Q_i + 1 \qquad (3.11)$$

Finally, according to [ITU98], after obtaining the P_i and Q_i coefficients, two ratios will be calculated, *i.e.*, $P_i + Q_i$ will be divided by the total number of scores given by each observer for the whole session, and $P_i - Q_i$ divided by $P_i + Q_i$ as an absolute value. The criterion to eliminate the observer i is given by [ITU98]:

$$\text{If } \frac{P_i + Q_i}{J \times K \times R} > 0.05 \text{ and } \left|\frac{P_i - Q_i}{P_i + Q_i}\right| < 0.3 \text{, the observer } i \text{ should be rejected} \qquad (3.12)$$

where,

N is the number of observers;

J is the number of test conditions for each video sequence;

K is the number of video test sequences;

R is the number of video sequences' repetitions during the session;

L is the number of test presentations (in most cases the number of presentations will be equal to $J \times K \times R$, however it is noted that some assessment may be conducted with unequal numbers of sequences for each test condition).

The observer elimination should not be applied more than once to the results of a given session [ITU99].

After the observer's validation has been performed the MOS computed previously and using (3.1), must be re-calculated. The "new" MOS is computed taking into account the number of observers N' which are in accordance with the observers validation criterion explained in (3.12). Thus, similarly to (3.1), the "new" MOS is given by,

$$\mathrm{MOS}' = \frac{1}{N'} \sum_{i=1}^{N} u_{ijkr} \qquad (3.13)$$

where, $N' = N - number\ of\ \ rejected\ observers$.

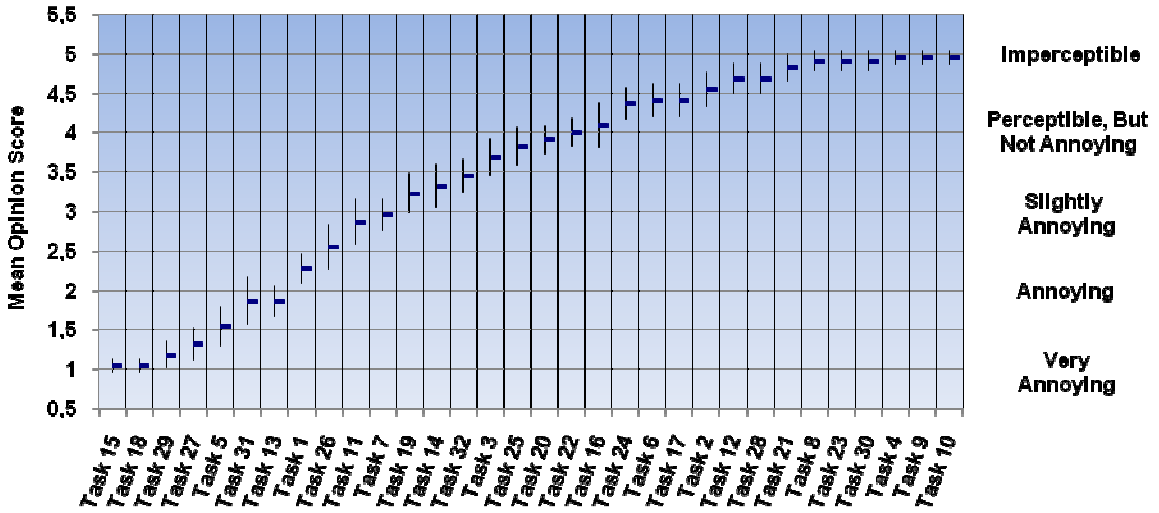
3.9 Subjective quality assessment results

In this section the results of the subjective quality assessment sessions are presented. Table 3.4 presents the MOS (computed according to 3.13) based in the opinion given by the observers in each one of the sessions, *i.e.*, using the video compression standards H.264 and MPEG-2. As it is possible to observe from Table 3.4, the values of the MOS show that the video compression applied to each video sequence required the observers to use the 5 grades of the scale. Figure 3.13 shows the MOS associated to each task for H.264 and MPEG-2 with the 95.5% confidence interval plotted as a vertical bar.

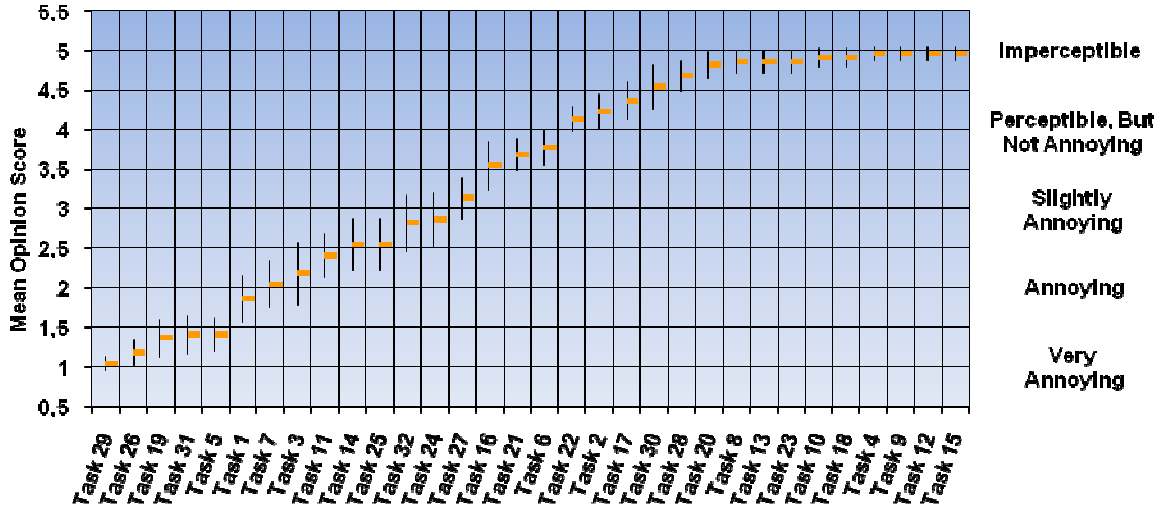
The video material used to perform the subjective tests, such as the original video sequences, the H.264 and MPEG-2 compressed videos, as well as the tests results – the opinion scores and the MOS – are available for those who are interested on the video quality evaluation field at http://amalia.img.lx.it.pt/~tgsb/H264_test/. Figure 3.14 shows some screenshots of the website where the test material used to perform the subjective tests is available.

Table 3.4: MOS using video compression standard H.264 and MPEG-2

	H.264			MPEG-2		
Task	Sequence	Rate [kbit/s]	MOS	Sequence	Rate [kbit/s]	MOS
Task 1	Australia	32	2.29	Container	128	1.86
Task 2	Table	256	4.57	Stephan	2048	4.24
Task 3	Container	64	3.71	Mobile	256	2.24
Task 4	Football	2048	4.95	Foreman	2048	4.95
Task 5	Mobile	128	1.57	Australia	128	1.43
Task 6	Coastguard	256	4.43	Table	512	3.81
Task 7	Foreman	128	3.00	Coastguard	256	2.05
Task 8	Stephan	1024	4.90	Football	2048	4.86
Task 9	Container	512	4.95	Mobile	4096	4.95
Task 10	Australia	256	4.95	Container	1024	4.90
Task 11	Table	128	2.95	Stephan	1024	2.43
Task 12	Mobile	512	4.71	Australia	1024	4.95
Task 13	Coastguard	64	1.90	Table	2048	4.86
Task 14	Football	512	3.38	Foreman	512	2.57
Task 15	Stephan	128	1.05	Football	4096	4.95
Task 16	Foreman	256	4.10	Coastguard	512	3.52
Task 17	Australia	128	4.38	Container	512	4.33
Task 18	Foreman	64	1.05	Coastguard	2048	4.90
Task 19	Coastguard	128	3.24	Table	256	1.38
Task 20	Mobile	256	3.95	Australia	512	4.81
Task 21	Container	256	4.86	Mobile	1024	3.67
Task 22	Football	1024	4.00	Foreman	1024	4.14
Task 23	Table	512	4.90	Stephan	4096	4.86
Task 24	Stephan	512	4.38	Football	1024	2.90
Task 25	Container	128	3.81	Mobile	512	2.57
Task 26	Stephan	256	2.57	Football	512	1.19
Task 27	Mobile	64	1.33	Australia	256	3.14
Task 28	Coastguard	512	4.71	Table	1024	4.67
Task 29	Table	64	1.19	Stephan	512	1.05
Task 30	Foreman	512	4.90	Coastguard	1024	4.57
Task 31	Football	256	1.90	Foreman	256	1.43
Task 32	Australia	64	3.48	Container	256	2.76



(a)



(b)

Figure 3.13: MOS with confidence interval of 95.5% for (a) H.264 and (b) MPEG-2

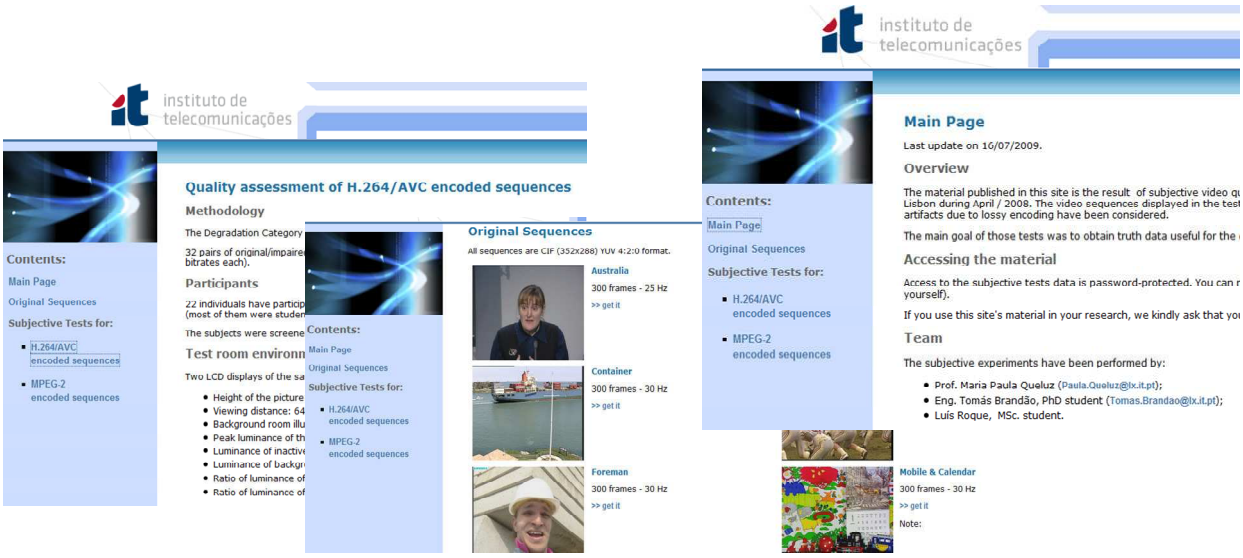


Figure 3.14: Website screenshots

Chapter 4

Objective Quality Evaluation

4.1 Introduction

As mentioned in the previous Chapter, subjective tests are particularly important in video quality evaluation since they provide the means to quantify quality as it is perceived by the viewers. However, they are not suitable for monitoring video data quality.

Recently, the increasing success of digital TV has motivated the research of objective quality evaluation metrics. These metrics aim to assess the quality of a broadcasted video as it is perceived at the user-end, automatically and in a real time basis. In order to validate the performance of an objective quality metric, the human assessment must also be taken into account.

This Chapter proposes two new objective video quality assessment metrics that combines a small set of features extracted from video sequences available at the user side. Regarding to the

objective video quality assessment metrics proposed in this chapter, a similar strategy it was followed by Tobias in [KOD09] and in [OD07], in which a NR and a RR video quality metrics were proposed, respectively. However the well succeeded results were not achieved for both video quality models. In fact, Tobias achieved much better results for the RR model proposed in [OD07] than for the NR model proposed in [KOD09].

In order to improve the metrics performance results a statistical technique, named as Principal Component Analysis (PCA), is used. This method is generally used for extracting the relevant information from a correlated feature data set transforming it into a smaller set of less correlated variables (called Principal Components).

This chapter is organized as follows. After the Introduction, in section 4.2 the MOS prediction model is proposed, based on a study of a set of video features and their effect on the MOS values. In the same section, PCA is described. In section 4.3 a set of measurements, proposed by VQEG, are presented with the intention of evaluating the MOS prediction model performance. In section 4.4, the model results and a performance evaluation of the proposed metrics for H.264 and MPEG-2 compression standards are presented. Finally, in section 4.5, the main conclusions resulting from the work reported in this chapter, are drawn.

4.2 Proposed MOS Prediction Algorithms

4.2.1 Motivations

The main goal of the proposed MOS prediction models is to estimate the quality value that a human observer would give to a video sequence. From the analysis of the subjective results it was possible to relate a set of video features with the correspondent MOS values. In choosing the video features that should be included in the MOS prediction model, a trade-off was set between the influence that each feature has on the MOS values and the difficulty of obtaining each one of them. Obviously not all the video sequence features influence the MOS in the same way, there are features which have a high impact in MOS values and others that can be discarded since they have not much influence on them.

In order to study the effect that some features have on the MOS values, an analysis is performed taking into account the individual effect of these features.

Figures 4.1 and 4.2 depict the MOS evolution with the bitrate and the MSE, respectively, for a set of video sequences displayed during the subjective tests.

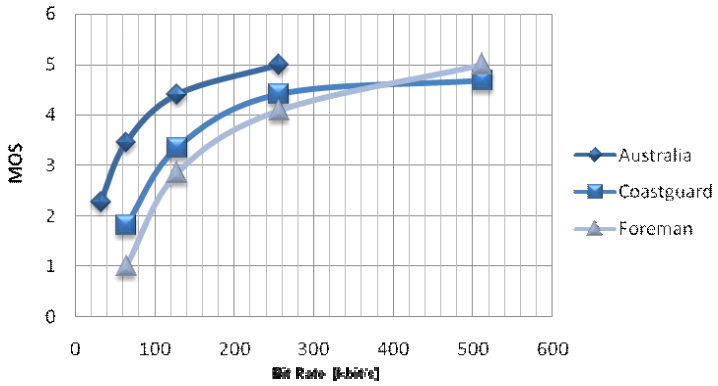


Figure 4.1: MOS evolution with bitrate of some video sequences

Figure 4.1 shows that, as expected, MOS increases with the bitrate of the encoded video sequences. Observing the figure, it can be seen that the MOS evolution with the bitrate is not linear: for high bitrates a large variation on the bitrate does not lead to a significant variation on the MOS; on the other hand, for low bitrates a small bitrate variation can conduct to a large MOS variation. The trend lines in Figure 4.1 can thus be described by a logarithmic function applied to the bitrate.

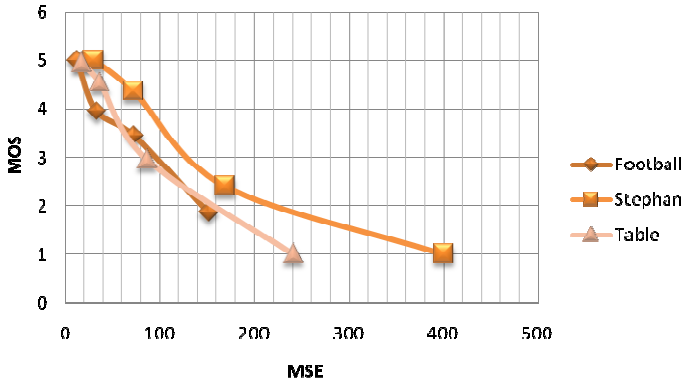
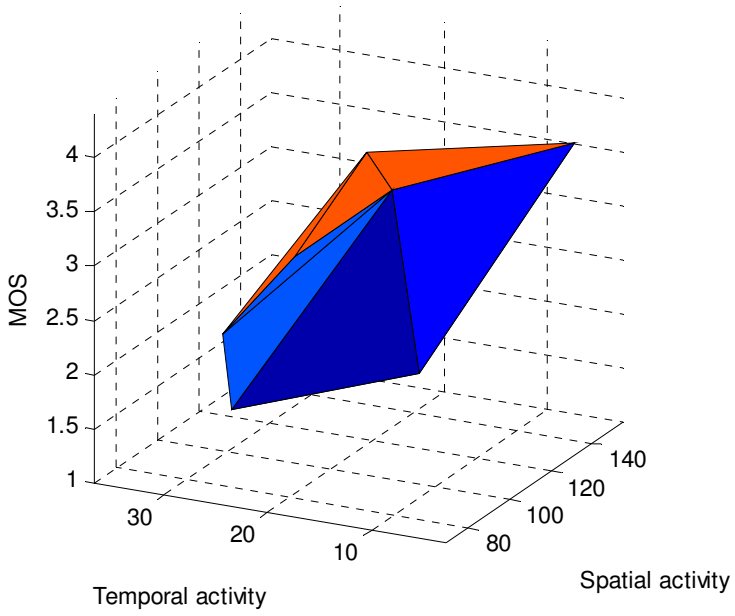


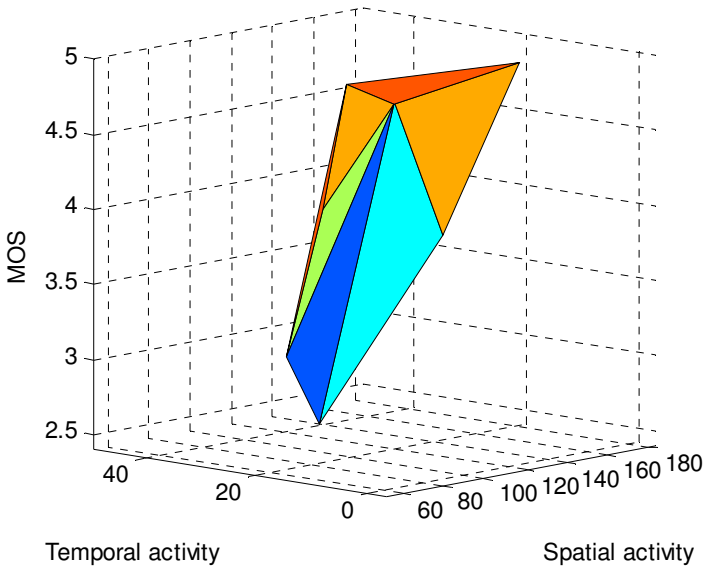
Figure 4.2: MOS evolution with the MSE of some video sequences

Figure 4.2 shows that there is also a relation between MOS and MSE. For this case, MOS values are roughly inversely proportional to the MSE values. Thus, the higher is the difference (MSE) between each frame of the original and the encoded video, the lower will be the grade given by the observers (MOS).

Other important observation taken from the subjective test results is the fact that the spatial activity and the temporal activity of the video sequences also influence the quality grades given by the observers. Figure 4.3 presents the resulting MOS values evolution with this spatial and the temporal activities for a set of video sequences encoded at two different bitrates.



(a)



(b)

Figure 4.3: MOS relation with spatial and temporal activities for a set of video sequences encoded at: (a) 128 kbit/s and (b) 1024 kbit/s

By observing Figures 4.3.a) and b), it is possible to state that, at the same bitrate, MOS values of video sequences with a large spatial-temporal activity are negatively affected when compared to video sequences with reduced spatial-temporal activity. In fact, the spatial-temporal activity of a video sequence has an important role in its video quality and, indirectly, in its video quality evaluation provided by the observers. The HVS is largely influenced by the movement and texture contents in the video. Thus, when a video sequence characterized by a large spatial-temporal activity is encoded at low bitrates, its quality is more affected than the videos which have reduced spatial-temporal activity.

In this thesis, the chosen features were those that immediately stood out as having a high impact in the overall MOS values. The selected features were:

- Bitrate (BR);
- Global Spatial Activity (gSA);
- Global Temporal Activity (gTA);
- Spatial Activity Variance (vSA);
- Temporal Activity Variance (vTA);
- Global MSE ($gMSE$);
- MSE Variance ($vMSE$).

Besides the global value (or mean value), the variance of the MSE and the activity features is also taken into account. This is important for the cases where the video sequence presents high variations of these features, since the global value may not be sufficient to characterize these features evolution along the sequence.

4.2.2 MOS prediction models

In this sub-section, two approaches for MOS prediction model are proposed. The objective is to develop MOS prediction models that are based only in Non-Reference (NR) features, *i.e.*, computed at the receiver side from the received video sequence. The first approach uses all the features listed at the end of section 4.2.1, except the MSE metric; the second approach, extends the first one by also considering the MSE feature. It is important to mention that this second approach has a higher computational complexity than the first one, resulting from the inclusion of an algorithm that estimates the MSE.

The features considered for the first model are:

- Bitrate (BR);
- Global Spatial Activity (gSA);
- Global Temporal Activity (gTA);

- Spatial Activity Variance (vSA);
- Temporal Activity Variance (vTA).

The low complexity model can be formally described by

$$\hat{MOS} = f_1(BR, gTA, gSA, vTA, vSA) \quad (4.1)$$

where $\hat{MOS}$ is the MOS prediction. In the second approach, all the features mentioned in section 4.2.1 are considered. Thus, this more complex MOS prediction model will be based on the following features:

- Bitrate (BR);
- Global Spatial Activity (gSA);
- Global Temporal Activity (gTA);
- Spatial Activity Variance (vSA);
- Temporal Activity Variance (vTA);
- Global MSE ($gMSE$);
- MSE Variance ($vMSE$).

This second MOS prediction model can be formally described by

$$\hat{MOS} = f_2(BR, gMSE, vMSE, gTA, gSA, vTA, vSA) \quad (4.2)$$

The MSE estimation is computed using the non-reference PSNR estimation algorithm developed by Brandão and Queluz [BQ08b]. An auxiliary method where the reference video is assumed to be known is proposed, in order to give an overview about the model functionality and, at same time, to validate the second approach model based only on the degraded video.

Although the inclusion of the MSE as a feature increases the system and computational complexity, it is of interest to evaluate its influence in the accuracy of the MOS estimation.

The PSNR estimation algorithm proposed in [BQ08b] explores statistical properties of the DCT coefficients, which are modelled by a Cauchy or Laplace probability density function. Table 4.1 depicts a comparison between the true PSNR and the estimated PSNR values computed by the algorithm, for all video sequences used in the subjective tests. The results show that the estimated PSNR is generally accurate, independently of the bitrate at which a video sequence is encoded.

The PSNR estimation model is therefore a robust and accurate alternative to the true PSNR value and, consequently, to the true MSE.

Table 4.1: True PSNR and the estimated PSNR values

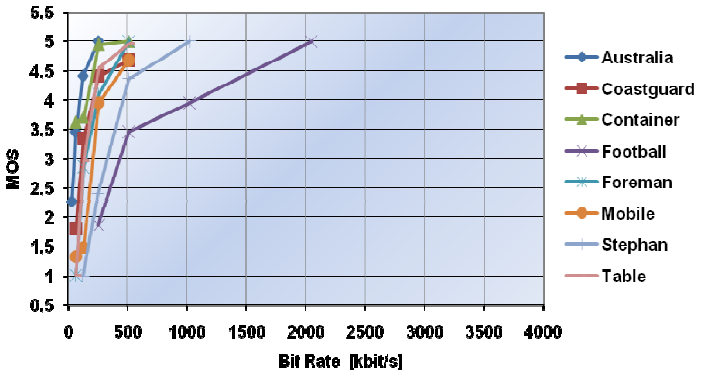
	H.264			MPEG-2		
Video sequence	Rate [kbit/s]	True PSNR	Estimated PSNR	Rate [kbit/s]	True PSNR	Estimated PSNR
Stephan	**128**	22,57	22,76	**512**	23,87	24,61
	256	26,47	26,15	**1024**	26,77	28,00
	512	30,26	29,42	**2048**	30,97	31,40
	1024	33,96	32,87	**4096**	36,05	36,33
Table	**64**	24,61	24,28	**256**	27,29	25,02
	128	29,19	28,51	**512**	31,20	32,08
	256	32,76	32,19	**1024**	35,24	36,14
	512	36,13	35,65	**2048**	39,15	39,95
Mobile	**64**	19,71	20,54	**128**	22,42	24,32
	128	20,81	20,61	**512**	23,73	26,94
	256	23,08	25,51	**1024**	26,54	28,78
	512	30,23	28,82	**4096**	34,71	35,02
Football	**256**	26,82	25,22	**256**	26,68	25,08
	512	30,03	28,44	**512**	26,73	25,13
	1024	33,52	32,08	**1024**	30,15	30,03
	2048	37,80	36,72	**2048**	34,54	34,17
Foreman	**64**	26,21	23,91	**128**	28,15	24,70
	128	30,15	28,61	**512**	30,97	31,11
	256	33,66	32,50	**1024**	34,66	35,62
	512	36,86	36,24	**2048**	38,13	38,74
Coastguard	**64**	25,33	25,29	**128**	26,28	24,92
	128	27,54	27,59	**256**	26,32	25,06
	256	29,81	29,69	**512**	29,41	30,29
	512	32,25	32,05	**1024**	32,40	33,21
Container	**64**	29,66	28,94	**128**	27,99	24,91
	128	32,94	32,32	**256**	29,45	29,20
	256	36,04	35,54	**512**	33,54	33,44
	512	39,14	38,79	**1024**	37,73	38,84
Australia	**32**	33,10	28,48	**64**	32,90	25,35
	64	36,95	32,23	**128**	37,83	34,90
	128	40,40	36,18	**512**	41,70	40,35
	256	43,02	39,95	**1024**	43,49	43,08

After presenting the concepts behind the proposed MOS prediction models it is necessary to perform an independent study of the evolution of MOS with each of the mentioned features. After this study, the MOS prediction model can be developed.

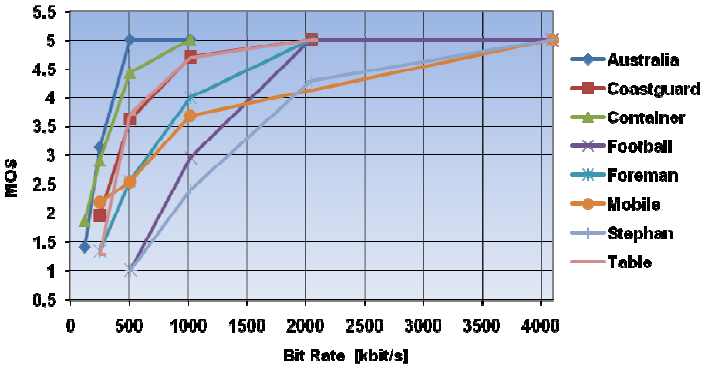
4.2.3 MOS evolution with each feature

This section presents an analysis of the MOS evolution with the features values. This analysis is important, since it will help to understand how these features contribute to the MOS values. The relation between each feature and the MOS will be modeled by a function in order to characterize

each feature influence on MOS. Figure 4.4.a) and Figure 4.4.b) depict the MOS evolution with the bitrate for video sequences encoded with H.264 and MPEG-2, respectively.



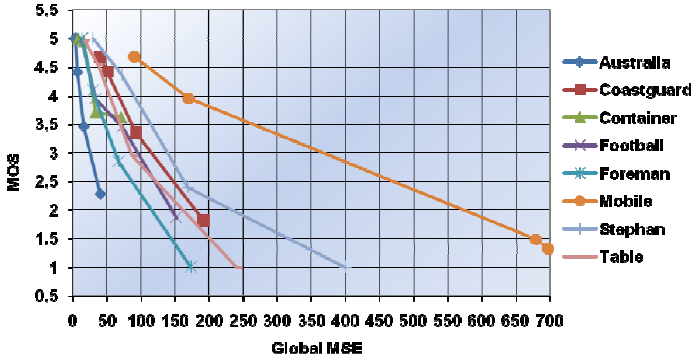
(a)



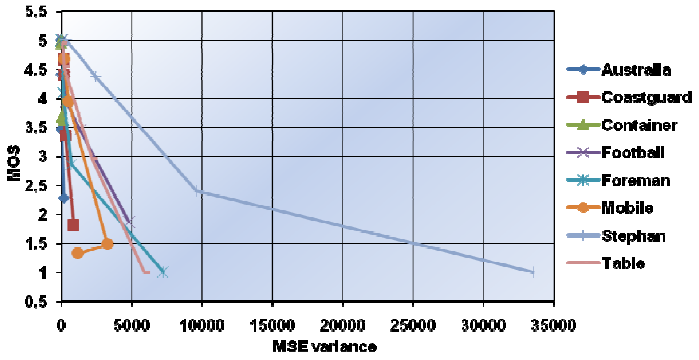
(b)

Figure 4.4: MOS evolution with the bitrate for a) H.264; b) MPEG-2

By observing the plots, it can be seen that, for both compression standards (H.264 and MPEG-2), the relation between the MOS and the Bitrate are similar, and can be roughly modeled by a logarithmic function. As for the MSE feature, its evolution with MOS is similar for both H.264 and MPEG-2. Figure 4.5 and Figure 4.6 present the MOS evolution with the global MSE and its variance, for H.264 and MPEG-2, respectively. Considering the evolution of these curves, a quadratic function has been used in order to “linearize” the MSE evolution with MOS.

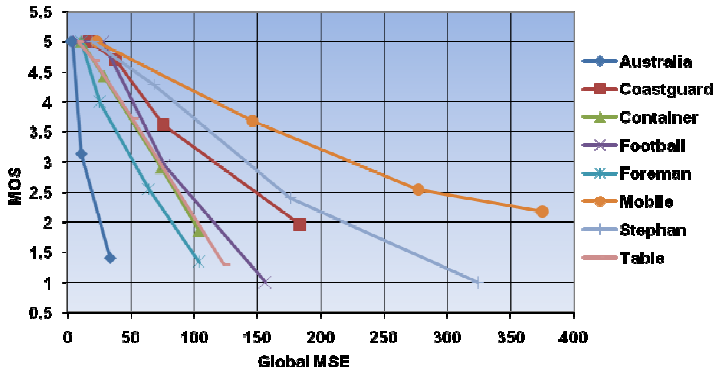


(a)

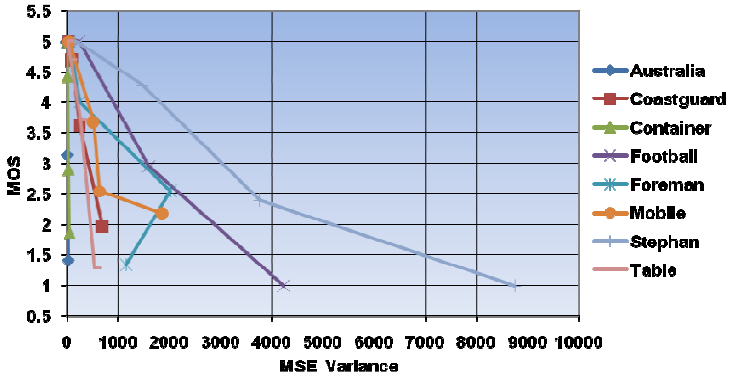


(b)

Figure 4.5: Relation between MOS and a) Global MSE for H.264; b) MSE Variance for H.264



(a)

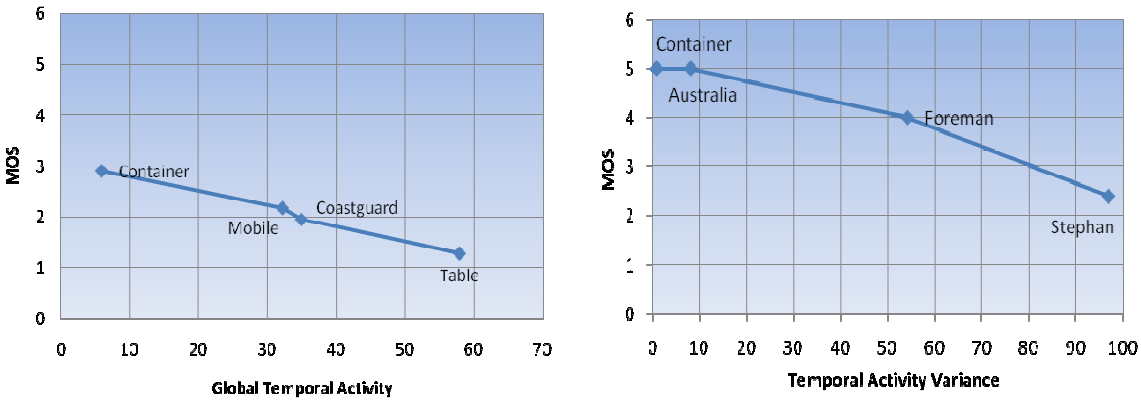


(b)

Figure 4.6: Relation between MOS and a) Global MSE for MPEG-2; b) MSE Variance for MPEG-2

In order to study the influence of the spatial and temporal activity on MOS values another strategy is followed. For the global temporal (or spatial) activity analysis, a set of video sequences with similar bitrate and global spatial (or temporal) activity have been selected. Analogously, for the temporal (or spatial) variance activity analysis only sequences with similarly bitrate and spatial (or temporal) variance activity should be considered.

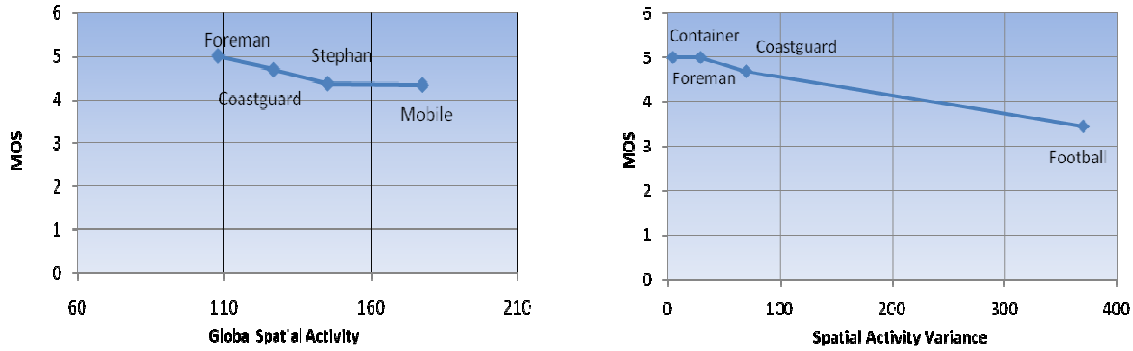
Figures 4.7.a) and b) show the MOS evolution with global temporal activity and temporal activity variance, for a fixed bitrate of 256 kbit/s (Figures 4.7.a)) and 1024 kbit/s (Figures 4.7.b)), respectively. From both figures, it is possible to see that the MOS values decrease linearly as the temporal activity increases.



(a) **(b)**

Figure 4.7: MOS evolution with: a) Global Temporal Activity; b) Temporal Activity Variance

Figure 4.8 shows the MOS evolution with the global spatial activity and spatial activity variance. Similarly to what happens for the temporal activity case, the relation between MOS and the global spatial activity (or the spatial activity variance) also shows a linear trend.



(a) **(b)**

Figure 4.8: MOS evolution with: a) Global Spatial Activity (512 kbit/s); b) Spatial Activity Variance (512 kbit/s)

After the analysis of the relation between feature values and MOS, it is necessary to combine them for accurately predict the MOS values. This combination will follow a linear regression model that is described in the following section.

4.2.4 Regression model

This sub-section describes the mathematical procedures for the proposed MOS prediction scheme, which are valid for the two approaches mentioned in section 4.2.2. Specifically, the predicted MOS value can be seen as the dependent variable in a linear equation, modelled as a function of the feature values and their corresponding linear weights. These weights are represented by $[\beta_0,...,\beta_n]$, where β_0 is the offset value and n represent the number of features.

Generically, this linear model is given by,

$$M\hat{O}S = \beta_0 + \sum_{i=1}^{n} \beta_i x_i \quad (4.3)$$

where $M\hat{O}S$ is the Mean Opinion Score prediction, n is the number of features, x_i is the value of feature i and $[\beta_0,...,\beta_n]$ are the linear weights, being β_0 the offset value.

However, not all the features presented in sub-section 4.2.2, such as the global MSE and the MSE variance, have a linear relation with the MOS. Thus, in order to compute these features contribution on the MOS prediction, individual regression models for those cases are proposed. In other words, in order to "linearize" the MSE evolution with the MOS, a quadratic function has been used before taking account all features contribution, presented in sub-section 4.2.2, on the final model.

Concerning the MOS evolution with the global MSE and the MSE variance, in sub-section 4.2.3 the quadratic relation of the MOS with both features was presented (Figure 4.5 and Figure 4.6). The partial MOS estimation, taking into account these features contribution, is given by,

$$\begin{aligned} M\hat{O}S_{gMSE} &= \beta_{0\,gMSE} + \beta_{1\,gMSE} \times gMSE + \beta_{2\,gMSE} \times gMSE^2 \\ M\hat{O}S_{vMSE} &= \beta_{0\,vMSE} + \beta_{1\,vMSE} \times vMSE + \beta_{2\,vMSE} \times vMSE^2 \end{aligned} \quad (4.5)$$

where,

$M\hat{O}S_{gMSE}$ and $M\hat{O}S_{vMSE}$	are the MOS estimation taking into account only one feature: the global MSE and the MSE variance, respectively;
$\beta_{0\,gMSE}$, $\beta_{1\,gMSE}$, $\beta_{2\,gMSE}$	are the linear weights that result from the $M\hat{O}S_{gMSE}$ regression;
$\beta_{0\,vMSE}$, $\beta_{1\,vMSE}$, $\beta_{2\,vMSE}$	are the linear weights that result from the $M\hat{O}S_{vMSE}$ regression;
$gMSE$ and $vMSE$	are the global MSE and the MSE variance, respectively.

The linear weights in (4.4) and (4.5) are computed using the features and MOS values in a set of training sequences assessed in the subjective tests.

A scheme of the MOS prediction model proposed in this thesis is depicted in Figure 4.9. As can be observed, the MOS prediction is computed through a linear regression that combines the selected features that influence MOS values.

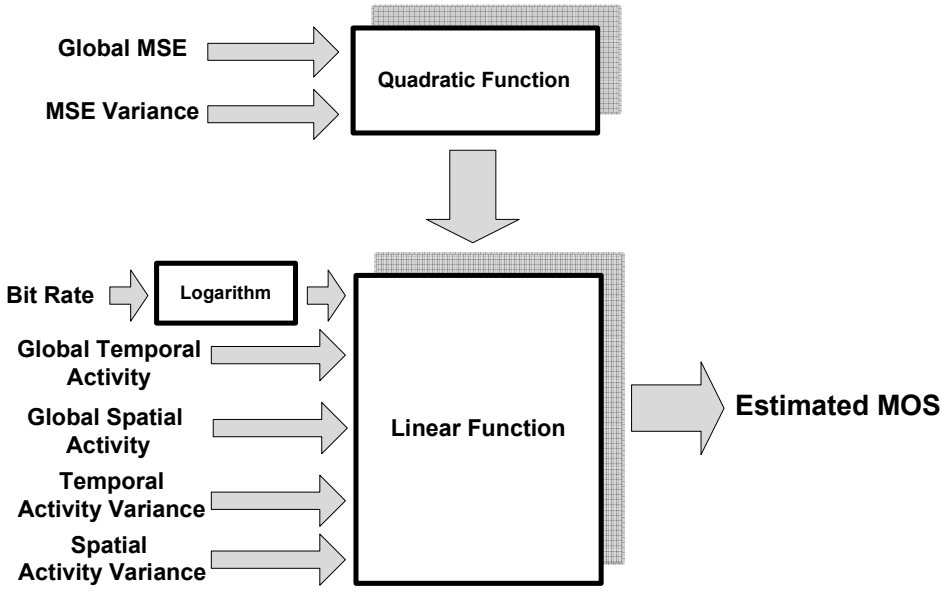


Figure 4.9: MOS prediction model description

As described in section 4.2.1, the MOS presents a linear evolution with logarithm of the bitrate. Thus, in order to obtain a better estimation of the MOS, the linear regression considers the logarithm of bitrate. Mathematically, the two linear models proposed in this thesis, based in high and low complexity systems, are given by (4.6) and (4.7), respectively.

$$\hat{MOS} = \beta_0 + \beta_1 \times \log(BR) + \beta_2 \times \hat{MOS}_{gMSE} + \beta_3 \times gSA + \beta_4 \times gTA + \beta_5 \times vSA + \\ + \beta_6 \times vTA + \beta_7 \times \hat{MOS}_{vMSE} \quad (4.6)$$

$$\hat{MOS} = \beta_0 + \beta_1 \times \log(BR) + \beta_2 \times gSA + \beta_3 \times gTA + \beta_4 \times vSA + \beta_5 \times vTA \quad (4.7)$$

The final question regarding the MOS prediction model design addresses the methodology for determining the linear weights. In order to determine adequate values for the weights, the importance of the subjective tests is highlighted once again since the MOS values from a set of training sequences will be used for determining those weights, β's.

Regarding the regression weights computation, one possible method to compute β is by minimizing the square error between MOS (the true MOS) and $\hat{MOS}$ (the estimated MOS), for the set of training video sequences. Since the training set is based on *K* video sequences with their corresponding MOS values, *K* feature vectors will be extracted for training. Thus, using the least square error criterion, the vector β is given by,

$$\begin{aligned}\hat{\beta} &= \arg\min_{\beta}\left\{\sum_{j=1}^{K}\left(MOS^{(j)} - M\hat{O}S^{(j)}\right)^2\right\} \\ &= \arg\min_{\beta}\left\{\sum_{j=1}^{K}\left[\left(\beta_0 + \sum_{i=1}^{N}\beta_i x_{ij}\right) - M\hat{O}S^{(j)}\right]^2\right\} \\ &= \arg\min_{\beta}\left\{\sum_{j=1}^{K}\left(x_j^T\beta - M\hat{O}S^{(j)}\right)^2\right\}\end{aligned} \tag{4.8}$$

which, in matrix form, is given by

$$\hat{\beta} = \arg\min_{\beta}\left\{\left[Y - X\beta\right]^T\left[Y - X\beta\right]\right\} \tag{4.9}$$

where,

$$X = \begin{bmatrix} 1 & x_1^{(1)} & \ldots & x_N^{(1)} \\ 1 & x_1^{(2)} & \ldots & x_N^{(2)} \\ \ldots & \ldots & \ldots & \ldots \\ 1 & x_1^{(K)} & \ldots & x_N^{(K)} \end{bmatrix}, \text{ and } Y = \begin{bmatrix} MOS^{(1)} \\ MOS^{(2)} \\ \ldots \\ MOS^{(K)} \end{bmatrix} \tag{4.10}$$

X is a $K \times N$ matrix, where each row contains the feature values taken from the *j*-th video sequence in the training set and Y is a vector with the true MOS values. Thus, the least squares solution for β can be computed according to,

$$\hat{\beta} = \left(X^T X\right)^{-1} X^T Y \tag{4.11}$$

After determining the weights, the best way to test their accuracy is to use a new set of video sequences (test sequences) and to compare the estimated MOS, computed by the prediction model, with the true MOS values taken from the subjective tests.

This control is even more interesting as different sequences are used, since the weights calibration is based solely on a limited number of sequences, and could deliver biased results.

4.2.5 Principal Component Analysis (PCA)

In this sub-section, in order to reduce and at the same time optimizing the model dimensionality without sacrificing the model accuracy, a purely mathematical method is conducted. The Principal Component Analysis, commonly known as PCA, is a technique which involves a mathematical procedure that transforms a number of correlated variables into a smaller number of uncorrelated variables called Principal Components. The PCA allows reducing the number of variables without losing the main information and consequently without losing the model's accuracy. This statistical technique quantifies the correlation between the features selected to predict the MOS and then, dependently of the amount of redundancy showed among each of other, they are combined in a

"new" feature. Thus, in order to measure the degree of the linear relationship between features, a covariance matrix is constructed. The elements of the covariance matrix, C_X, are given by,

$$c_{ij} = E\{ (x_i - \mu_i)(x_j - \mu_j)^T \} \quad (4.8)$$

where $(x_i - \mu_i)$ and $(x_j - \mu_j)$ denote the difference between the values of two features and its correspondent mean values, respectively. The elements of the matrix C_X, denoted by c_{ij}, represent the covariance between the random variables x_i and x_j; the elements c_{ii} represent the variance of the variable x_i. If two variables x_i and x_j are uncorrelated, their covariance is zero ($c_{ij} = c_{ji} = 0$). The covariance matrix is always symmetric, by definition.

In order to know which are the most significant components of the data set, it is necessary to compute an orthogonal basis by finding its eigenvalues and eigenvectors. The eigenvectors and the corresponding eigenvalues are the solutions of the equation,

$$C_X e_i = \lambda_i e_i \quad , \; i = 1,...,n \quad (4.9)$$

where λ_i are the eigenvalues and e_i are the correspondent eigenvectors. The eigenvectors correspond to the principal components of the data set, while the eigenvalues are their variance. The eigenvalues $\{\lambda_i$,where, $i = 1,...,n\}$ indicate the relative contribution of the transformed vectors, also named as principal components, to the total energy of the vector X. In fact, it should be mentioned that it is based on the eigenvalues, λ_i, that the principal components are selected, *i.e.*, among the eigenvectors set, the ones that have the highest eigenvalues are the ones that are the principal components of the data set.

The matrix of eigenvectors will diagonalize the covariance matrix C_X. After this analysis, the eigenfactors are less correlated and present less redundancy between each other. Therefore the MOS prediction model will be more effective and optimized.

4.3 Metrics Performance

In order to validate an objective quality metric, *i.e.*, to evaluate how well the objective model predicts the subjective judgements, it is necessary to quantify the performance of the model. Thus, a

set of measurements, proposed by VQEG[8], are described in this section. These measurements are the prediction accuracy, the prediction monotonicity and the prediction consistency. Additionally, it is suggested the computation of the Root Mean Square error (RMS error).

The set of statistical measurements that validate the performance of an objective quality assessment metric are the following:

- **Pearson Coefficient (prediction accuracy measurement)**

This statistical measurement is widely used for measuring the correlation between two variables. The Pearson Coefficient is a value that represents the strength of the linear relation between two variables. In the ideal situation, the correlation between subjective and objective MOS values would be equal to one. However, this ideal situation is very difficult to achieve, therefore it is fair to expect a Pearson Coefficient ranged between 0.9 and 1. Lower values for the Pearson Coefficient usually mean that the objective metric used in the MOS estimation is not an adequate one. The Pearson Coefficient is obtained from the expression (4.10).

$$P_c = \frac{\left(\sum_{i=1}^{N} x_i \times y_i\right) - \left(\frac{1}{N} \times \sum_{i=1}^{N} x_i \times \sum_{i=1}^{N} y_i\right)}{\sqrt{\left[\left(\sum_{i=1}^{N} {x_i}^2\right) - \frac{1}{N}\left(\sum_{i=1}^{N} x_i\right)^2\right] \times \left[\left(\sum_{i=1}^{N} {y_i}^2\right) - \frac{1}{N}\left(\sum_{i=1}^{N} y_i\right)^2\right]}} \qquad (4.10)$$

where N is the total number video sequences under evaluation, x_i is the MOS achieved in the subjective video quality tests and y_i is the predicted MOS by using an objective metric.

- **Spearman Coefficient (prediction monotonicity measurement)**

The Spearman Coefficient evaluates the degree to which the model's predictions agree with the relative magnitudes of the subjective quality scores. This coefficient evaluates how well a monotonic function could represent the relation between the two variables. Similarly to Pearson's Coefficient, an acceptable objective metric should lead to a value for the Spearman Coefficient ranging between 0.9 and 1.

The Spearman Coefficient is given by:

$$S_c = 1 - \left[\frac{6}{N \times \left(N^2 - 1\right)} \sum_{i=1}^{N} {d_0}^2 \left(x_i, y_i\right)\right] \qquad (4.11)$$

[8] http://www.its.bldrdoc.gov/vqeg

with,

$$d_0(x_i, y_i) = rank(x_i) - rank(y_i) \tag{4.12}$$

where x_i and y_i are the "true" MOS and the predicted MOS values for sequence i, respectively. $rank(x_i)$ and $rank(y_i)$ represent are the positions that each variable x_i and y_i assume in their sorted list of values.

- **Outlier Ratio (prediction consistency measurement)**

The Outlier Ratio is a statistical measurement that evaluates how well does the model maintain accuracy over the test range. The Outlier Ratio is given by:

$$Outlier\ \ Ratio = \frac{N_0}{N} \tag{4.13}$$

where N_0 is the number of outlier points and N is the total number of data points. The outlier points are all points of y_i that fall outside the interval given by $[x_i - 2\times\sigma_i, x_i + 2\times\sigma_i]$, where σ_i is the standard deviation of the opinion scores given for sequence i and x_i is the MOS achieved in the subjective video quality tests.

- **Root Mean Square error (RMS error)**

In this context, the RMS error measures the amount by which the estimated values of MOS differ from their "true" values. The higher is the gap between predicted and "true" MOS values the higher is the RMS error since it uses a quadratic difference between the estimated value of MOS and the "true" MOS values (4.14).

A small value of the RMS error is not enough to consider the objective metric as an adequate one. However, a large RMS error value may be a strong evidence that the objective metric is an inadequate one.

The RMS error is given by:

$$RMS = \sqrt{\sum_{i}^{N} \frac{(x_i - y_i)^2}{N}}, \tag{4.14}$$

where N is the total number video sequences used in the subjective tests, x_i is the MOS achieved in the subjective video quality tests and y_i is the predicted MOS by using an objective metric.

4.4 Results and parameters analysis

In this section, the results obtained with the implementation of the two approaches, described in the previous section, are presented for both H.264 and MPEG-2 encoding standards.

Two sets of data are used for training and testing. The training set is employed for model calibration, while the test set is used for evaluating the model's accuracy. It is empirically a good practice to use 1/3 of the available samples for training and the remaining 2/3 for testing. Since there are 32 video sequences available from the subjective tests and in order to have a more detailed analysis regarding the influence of the number of training sequences in the MOS accuracy, 3 configurations of training/test sequences have been considered:

- 12 sequences for training and 20 for testing;
- 15 sequences for training and 17 for testing;
- 18 sequences for training and 14 for testing.

Note that, in order to perform this study, the training sets should be selected as follows:

- the first set should be compound of 12 randomly chosen video sequences from the 32 available;
- the second set, of size 15, should consist on the 12 video sequences of the first set plus three other randomly chosen video sequences from the 20 available;
- the last set, of size 18, should be compounded by the second set plus three other randomly chosen video sequences from the 17 available.

In what concerns to the training and testing sets, there were no additional criteria for selecting the video sequences for each set. Since prediction accuracy and monotonicity are important topics concerning the legitimacy of the model, in order to perform a deeper analysis regarding the model's performance, tests were conducted and the results are presented throughout quantitative indicators.

4.4.1 Low complexity model

The first tested MOS prediction model is the low complexity model, based on the smaller set of features detailed in sub-section 4.2.2 and given by equation (4.7). These regression weights were estimated based on the values of MOS acquired in the subjective tests, and on the values of features extracted from the sequences in the training set.

Table 4.2 shows the regression weights values for H.264 and MPEG-2, respectively, using the three configurations of training/test sequences.

Table 4.2: Regression weights for the low complexity model: (a) for H.264 and (b) MPEG-2

(a)

Training/Test	β_0	β_1	β_2	β_3	β_4	β_5
12/20	3,8939	1,2850	-0,0135	-0,9428	-0,0451	0,5595
15 /17	3,5303	1,4803	0,0262	-1,3453	-0,0451	0,6929
18 /14	3,4722	1,5379	-0,0372	-1,3905	-0,0376	0,6905

(b)

Training/Test	β_0	β_1	β_2	β_3	β_4	β_5
12/20	3,5455	1,8124	-0,0417	-1,1918	0,4354	0,3351
15 /17	3,6606	1,6760	-0,0071	-1,0196	0,2936	0,2866
18 /14	3,7626	1,5726	0,0117	-1,0483	0,2718	0,2755

In Table 4.2.(a) and (b), besides β_0 (the offset value) it is possible to distinguish, based on their larger absolute values, two regression weights that stand out from both tables: β_1 and β_3. Regarding β_0, it is directly related with MOS mean of all video sequences comprised on the training set used to calibrate the regression parameters. Thus, it is an offset value. Relatively to parameters β_1 and β_3, since the feature values are normalized, it is possible to analyze the relevance that each corresponding feature has on MOS prediction. Thus, in accordance with the values that β_1 and β_3 take on Table 4.2.a) and Table 4.2.b), it is possible to observe that they have the larger values in absolute, meaning that the feature related with β_1, the bitrate, and with β_3, the global temporal activity, have a higher effect on MOS prediction comparatively with the remaining features. In fact, the feature related with β_3, the global temporal activity, contributes negatively on the MOS value. In this case, according to the linear regression (4.7), the global temporal activity, on contrary to bitrate, vary inversely with MOS values.

After estimating these parameters, the function (4.7) is applied to compute the $\hat{MOS}$ from the features set of values obtained from the video sequences of the testing set. Figures 4.10 and 4.11 display the results of the estimated MOS vs “true” MOS for H.264 and MPEG-2, respectively.

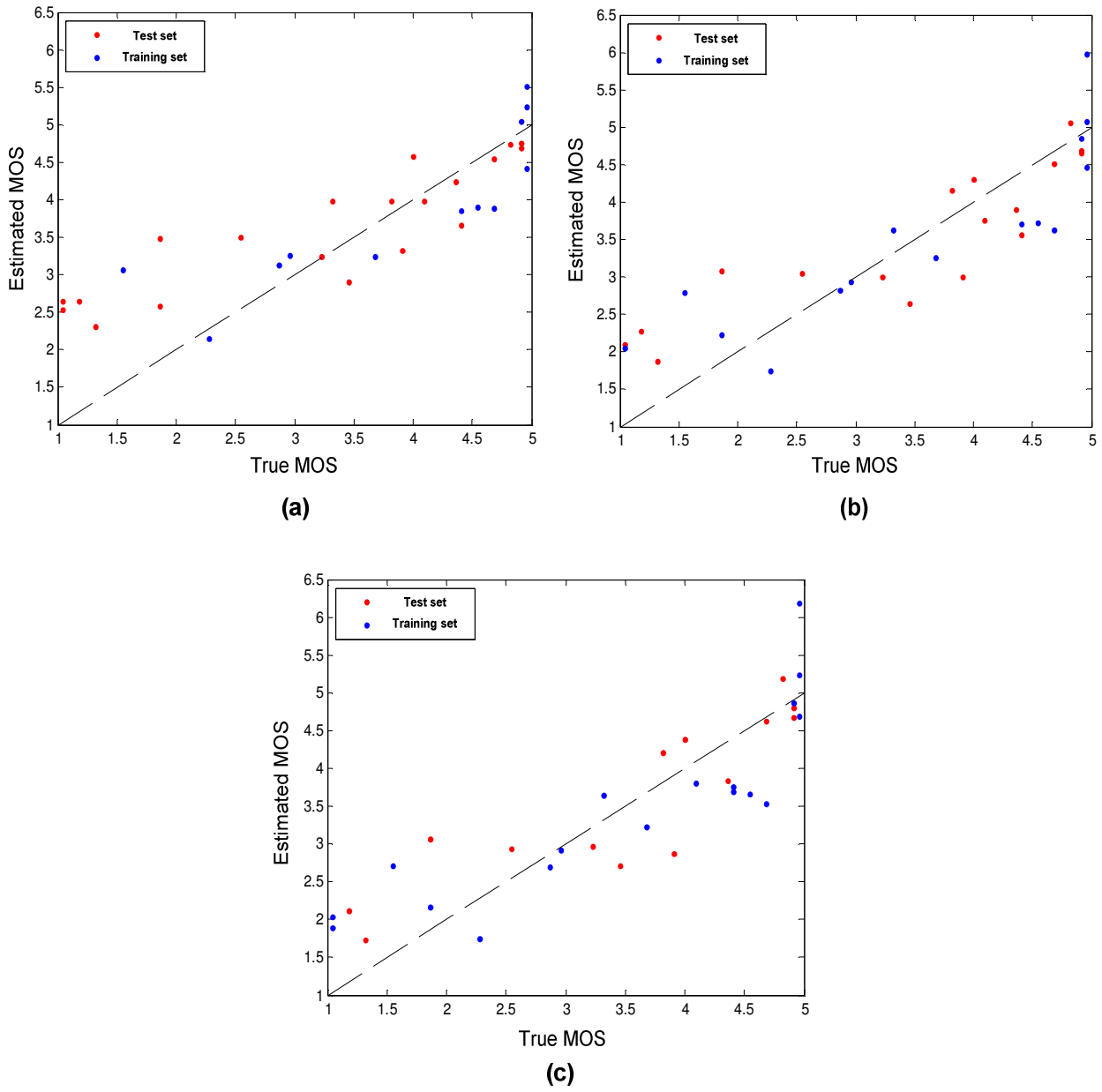


Figure 4.10: MOS estimation result for H.264: (a) 12 training/20 test video sequences; (b) 15 training/17 test video sequences; (c) 18 training/14 test video sequences

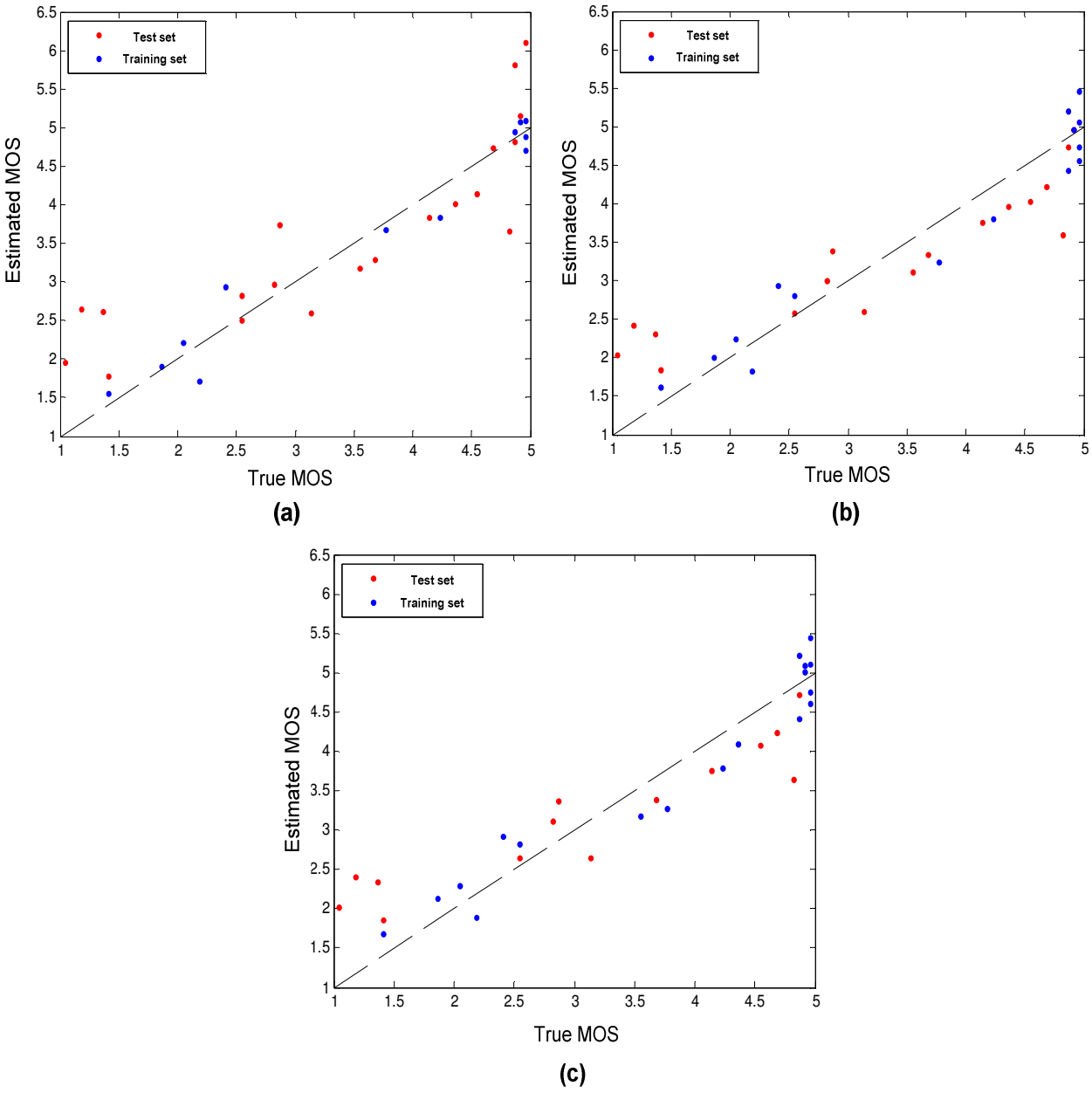


Figure 4.11: MOS estimation result for MPEG-2: (a) 12 training/20 test video sequences; (b) 15 training/17 test video sequences; (c) 18 training/14 test video sequences

To assess the resulting error, performance metrics, as described in section 4.3, are used. Table 4.3 depicts the results for Root Mean Square (RMS), Outlier Ratio, as well as the Pearson (P_c) and Spearman (S_c) coefficients, for H.264 and MPEG-2 compressed video.

Table 4.3: Metrics performance for: (a) H.264 and (b) MPEG-2

(a)

	Training				Test				Global (Training+Test)			
Training/Test	**RMS**	**Outlier Ratio**	P_c	S_c	**RMS**	**Outlier Ratio**	P_c	S_c	**RMS**	**Outlier Ratio**	P_c	S_c
12/20	**0,6280**	**0,2500**	**0,8375**	**0,9371**	**0,8414**	**0,2500**	**0,8811**	**0,8872**	**0,7683**	**0,2500**	**0,8545**	**0,8960**
15 /17	**0,6746**	**0,3333**	**0,8613**	**0,9250**	**0,6578**	**0,1765**	**0,8817**	**0,8848**	**0,6657**	**0,2500**	**0,8675**	**0,9162**
18 /14	**0,6898**	**0,2222**	**0,8633**	**0,9030**	**0,6031**	**0,1429**	**0,8814**	**0,8681**	**0,6533**	**0,1875**	**0,8697**	**0,9209**

(b)

	Training				Test				Global (Training+Test)			
Training/Test	**RMS**	**Outlier Ratio**	P_c	S_c	**RMS**	**Outlier Ratio**	P_c	S_c	**RMS**	**Outlier Ratio**	P_c	S_c
12/20	**0,2650**	**0,0000**	**0,9814**	**0,8881**	**0,7132**	**0,3000**	**0,8622**	**0,9083**	**0,5867**	**0,1875**	**0,9073**	**0,9190**
15 /17	**0,3493**	**0,0667**	**0,9660**	**0,9250**	**0,6313**	**0,1765**	**0,9192**	**0,9387**	**0,5185**	**0,1250**	**0,9307**	**0,9368**
18 /14	**0,3412**	**0,0556**	**0,9635**	**0,9319**	**0,6664**	**0,2143**	**0,9320**	**0,9297**	**0,5097**	**0,1250**	**0,9345**	**0,9416**

As expected, the performance of the estimation for the training set is better than for the testing set. Such result arises from the fact that the model's calibration is based on the training set for which is optimised. When testing the model with different video sequences, the performance decreases.

Regarding to the performance indicators values used for measuring the MOS prediction accuracy and reliability, the model presents, in general, lower results for Outlier Ratio and RMS with video sequences compressed with the MPEG-2 compression standard than for H.264 compressed sequences; the same trend can be observed for Pearson and Spearman coefficients.

In short, taking into account the individual analysis of each compression standard performance, it is possible to conclude that when MOS prediction model uses video sequences compressed with MPEG-2, the model gives a better MOS estimation than when the model uses video sequences encoded with the H.264 standard.

4.4.2 High complexity model

In this sub-section, the second MOS prediction model is described. It is characterized as a high complexity based model, since besides all features considered on 4.4.1, it includes an additional metric, the MSE. The computation of this feature will add an extra complexity to this approach, resulting from the inclusion of the algorithm that estimates the MSE. The MSE estimation is computed by using the PSNR estimation model developed by Brandão and Queluz [BQ08b]. Although there is an increase on the system complexity, it is of interest to evaluate the influence of this feature on the accuracy of the MOS estimation. In order to have a first sight about the model functionality, an auxiliary method, where the reference video is assumed to be known, was used. With regard to this auxiliary method, the MSE is computed between each correspondent frame from the degraded and

the reference video (true MSE). Mathematically, the MOS prediction model taken in this sub-section is given by (4.8). Analogously to the first model, described in section 4.4.1, the influence of the size of the training set on the MOS estimation was studied using three training sets of size 12, 15 and 18 video sequences, to calibrate the regression weights, $[\beta_0,...,\beta_7]$. These regression weights were also estimated taking into account the same MOS values presented in section 3.9, and the objective metric's set of values comprised on the training set. The analysis of the results will be performed independently for the H.264 and MPEG-2 compression standards.

4.4.2.1 Results Analysis for the H.264 compression standard

In Table 4.4, the regression weights values for H.264 using the "true" MSE and the estimated MSE, respectively, for the three configurations of training/test sequences are presented.

Table 4.4 : Regression weights for the high complexity model taking into account H.264 compressed video sequences using: (a) the "true" MSE; (b) the estimated MSE

(a)

Training/Test	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7
12 /20	3.7311	0.7195	0.4717	0.6758	-0.4970	-0.7091	1.2057	-0.2789
15 /17	3.4000	0.6853	0.4476	0.6627	-0.5170	-0.6949	1.1646	-0.1401
18 /14	3.3636	0.6483	0.4260	0.7735	-0.5556	-0.7624	1.2584	-0.1942

(b)

Training/Test	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7
12 /20	2.8826	0.3126	0.1848	0.3608	-0.2556	-0.1719	1.1987	0.0292
15 /17	2.9333	0.4386	0.0588	0.2549	-0.1880	-0.1170	1.0271	0.0657
18 /14	3.2197	0.4676	0.0650	0.2097	-0.1968	-0.0700	0.9737	0.0833

Similarly to the low complexity based model described in sub-section 4.4.1, the offset value, β_0, highlights from the other parameters since it represents the MOS mean of all video sequences included on the training set. Besides β_0, it is possible to distinguish two regression parameters that stand out from both tables (β_1 and β_6) due to their absolute value higher than the remaining weights. From equation (4.6), β_1 and β_6 are associated to the bitrate logarithm and the temporal activity variance, respectively, meaning that these features are the ones that have more impact in the $\hat{MOS}$ value.

The features related with β_4, the global temporal activity and β_5, the spatial activity variance contributes negatively on the predicted MOS value, *i.e.*, the $\hat{MOS}$ will decrease when the global

temporal activity or the spatial activity variance increases. After estimating these parameters, the function (4.6) is applied to compute the $\hat{MOS}$ from the features set of values obtained from the video sequences of the testing set. Figures 4.12 and 4.13 display the MOS prediction results using the "true" MSE and the estimated MSE, respectively.

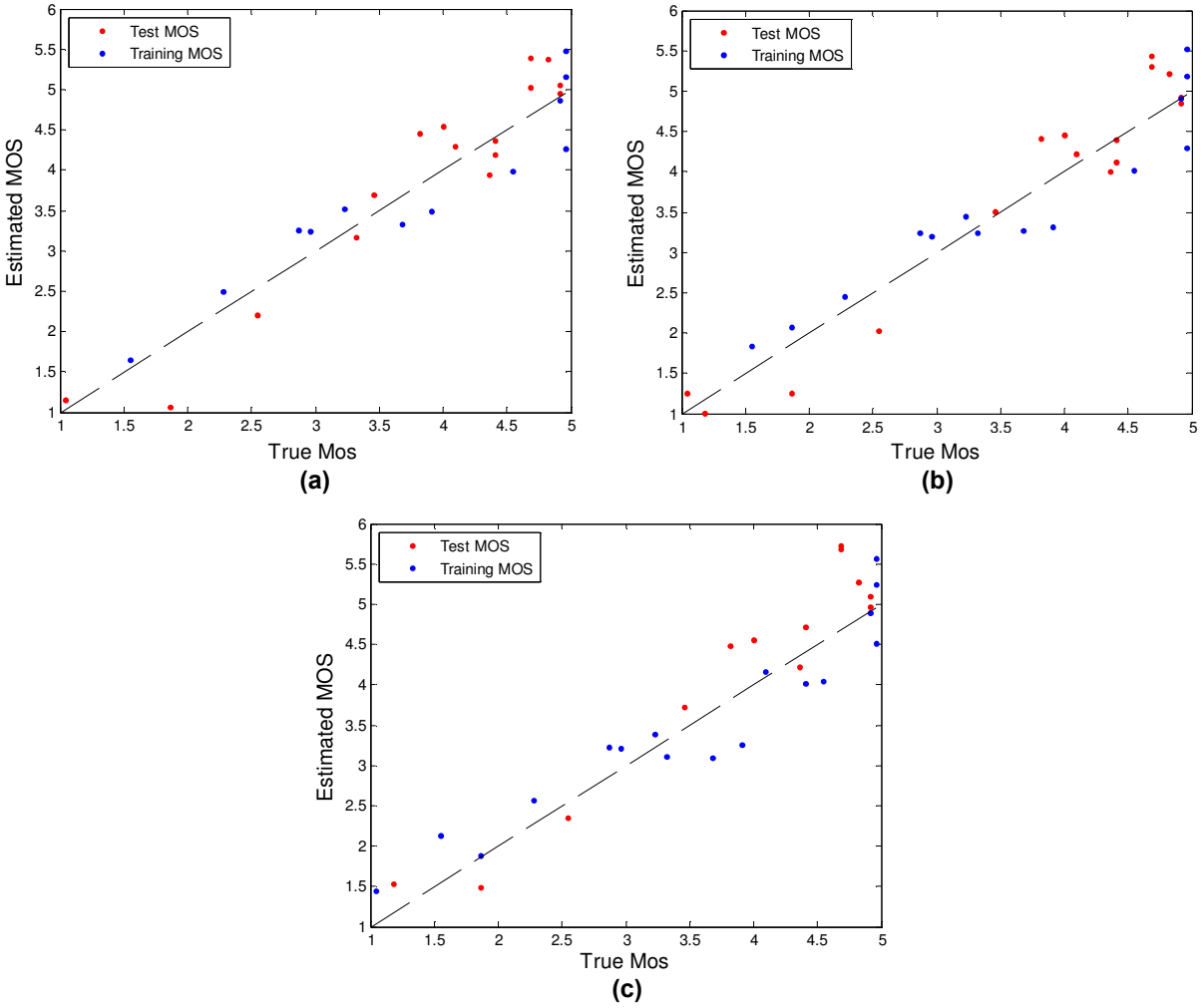


Figure 4.12: MOS estimation result for H.264 using the "true" MSE: (a) 12 training/20 test video sequences; (b) 15 training/17 test video sequences; (c) 18 training/14 test video sequences

As can be seen, according to Figure 4.12 and Figure 4.13, the MOS prediction model, independently on how the MSE was computed shows to be more accurate when the number of training sequences increases.

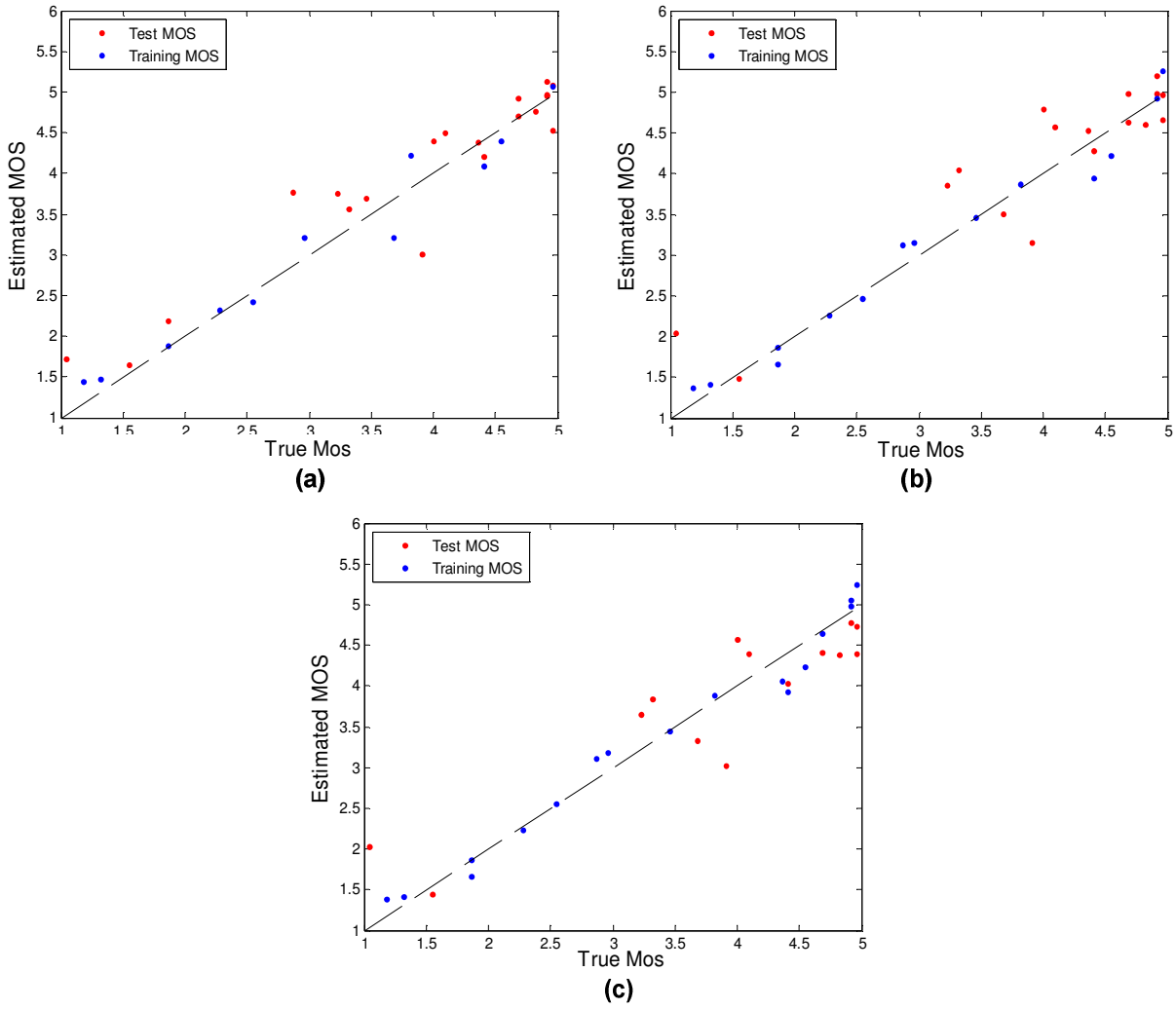


Figure 4.13: MOS estimation result for H.264 using the estimated MSE: (a) 12 training/20 test video sequences; (b) 15 training/17 test video sequences; (c) 18 training/14 test video sequences

Although this graphic analysis provides a general overview about the model performance, this kind of analysis is not conclusive about the model performance.

Since prediction accuracy and monotonicity are important features concerning the legitimacy of the model and are not possible to estimate through simple observation, the performance metrics described in section 4.3 are computed for a proper performance study.

Tables 4.5.a) and 4.5.b) depict the results obtained from the calculation of the indicators RMS, the Outlier Ratio, the P_c and the S_c, using the true MSE as well as the predicted MSE.

Table 4.5: Model performance analysis for H.264 using: (a) the "true" MSE; (b) the estimated MSE

(a)

	Training				Test				Global (Training+Test)			
Training/Test	**RMS**	**Outlier Ratio**	P_c	S_c	**RMS**	**Outlier Ratio**	P_c	S_c	**RMS**	**Outlier Ratio**	P_c	S_c
12 /20	0.3848	0.1667	0.9393	0.9300	0.7144	0.1500	0.9717	0.9173	0.5120	0.1563	0.9535	0.9212
15 /17	0.3688	0.1333	0.9571	0.9536	0.4074	0.0000	0.9742	0.8775	0.3898	0.0625	0.9631	0.9260
18 /14	0.3911	0.1111	0.9550	0.9381	0.5273	0.1429	0.9719	0.8901	0.3557	0.0250	0.9653	0.9326

(b)

	Training				Test				Global (Training+Test)			
Training/Test	**RMS**	**Outlier Ratio**	P_c	S_c	**RMS**	**Outlier Ratio**	P_c	S_c	**RMS**	**Outlier Ratio**	P_c	S_c
12 /20	0.2489	0.0000	0.9823	0.9930	0.4065	0.1000	0.9484	0.9143	0.4057	0.0625	0.9557	0.9409
15 /17	0.2034	0.0000	0.9880	1.0000	0.4699	0.0588	0.9197	0.8554	0.3697	0.0313	0.9627	0.9568
18 /14	0.2038	0.0000	0.9888	0.9959	0.5048	0.2143	0.9075	0.8154	0.3672	0.0438	0.9610	0.9521

The first conclusion that can be drawn from Tables 4.5.a) and 4.5.b) is that, independently of which MSE the MOS prediction model is using (true or estimated), the metrics performance results are similar. Considering these results, it is possible to support the algorithm's legitimacy, independently if the model uses the true MSE or the estimated MSE provided by [BQ08b]

Additionally, based on the values taken by RMS, Outlier Ratio, P_c and S_c, it is possible to notice a clear improvement trend on the values of those performance metrics when the number of training sequences increases.

4.4.2.2 Results Analysis for the MPEG-2 compression standard

In Tables 4.6.a) and b), the regression weights values for MPEG-2 using the "true" MSE and the estimated MSE, respectively, are presented.

Table 4.6: Regression weights for the three training/test configurations for MPEG-2 using: (a) the "true" MSE; (b) the estimated MSE

(a)

Training/Test	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7
12 /20	3.5455	1.0576	0.1321	-0.4564	0.2345	0.3104	0.2011	0.3923
15 /17	3.6606	0.8612	0.1252	-0.1817	0.0902	0.0748	0.4970	0.1261
18 /14	3.7626	0.7905	0.1474	-0.2057	0.0708	0.0513	0.4778	0.1374

(b)

Training/Test	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7
12 /20	3.5455	0.4114	0.0414	-0.0294	0.0585	-0.0729	1.1030	-0.0335
15 /17	3.6606	0.3546	0.0226	0.0019	-0.0459	-0.0733	0.8236	0.3000
18 /14	3.7626	0.3295	0.0362	-0.0415	-0.0307	-0.0345	0.8022	0.2474

According to table 4.6.a) and b), due to the large absolute value of some regression weights, beyond the β_0 parameter, the offset value, it is possible to distinguish one more regression weight that highlight from each one of the tables. These regression weights are respectively, β_1 from Table 4.6.a) and β_6 from Table 4.6.b). From equation (4.6), described in sub-section 4.2.4, β_1 and β_6 are related to the bitrate logarithm and the temporal activity variance, respectively, meaning that these features are the ones that have a higher contribution on the $\hat{MOS}$ value than the remaining. Figure 4.14 and Figure 4.15 presents graphically the MOS prediction results for the three training/test sequences configurations using the "true" MSE and the estimated MSE, respectively.

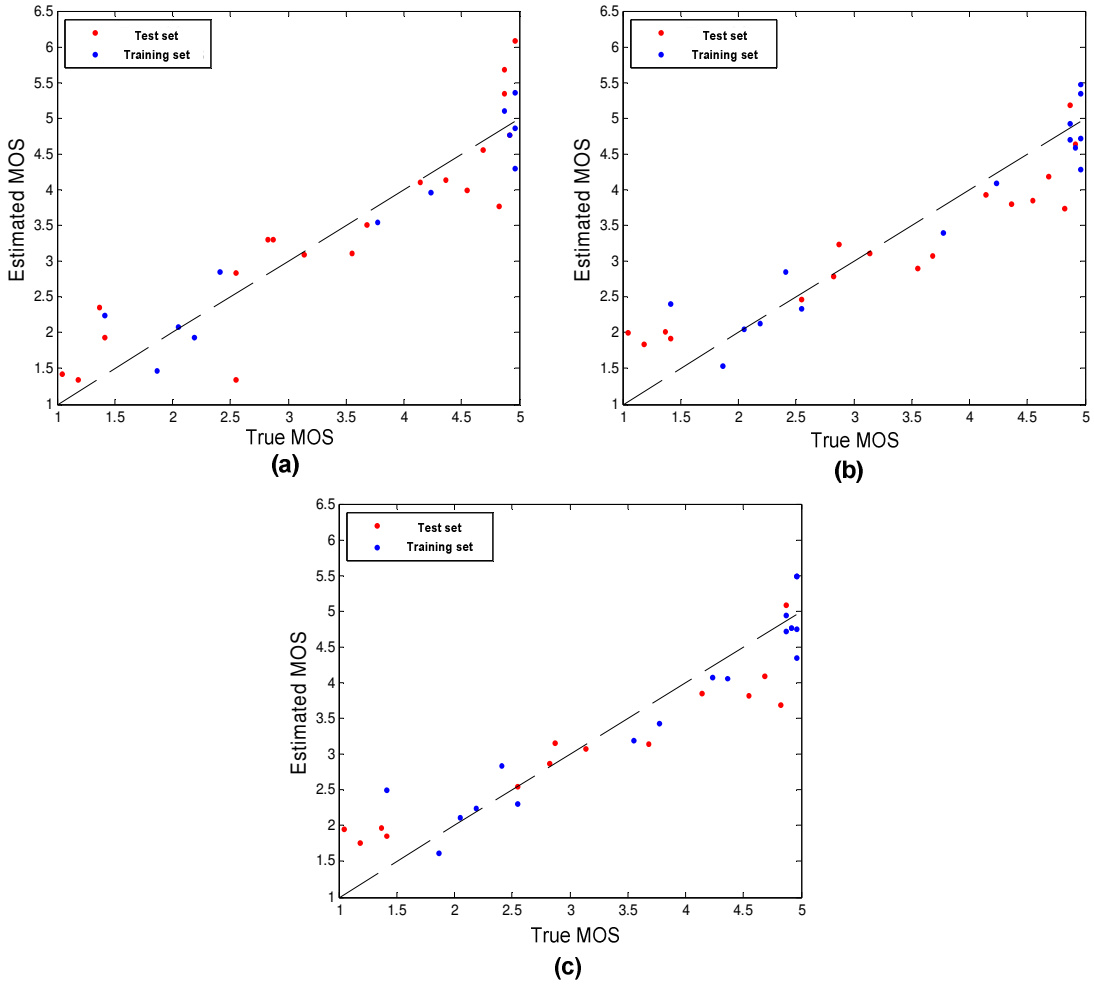


Figure 4.14: MOS estimation result for MPEG-2 using the "true" MSE: (a) 12 training/20 test video sequences; (b) 15 training/17 test video sequences; (c) 18 training/14 test video sequences

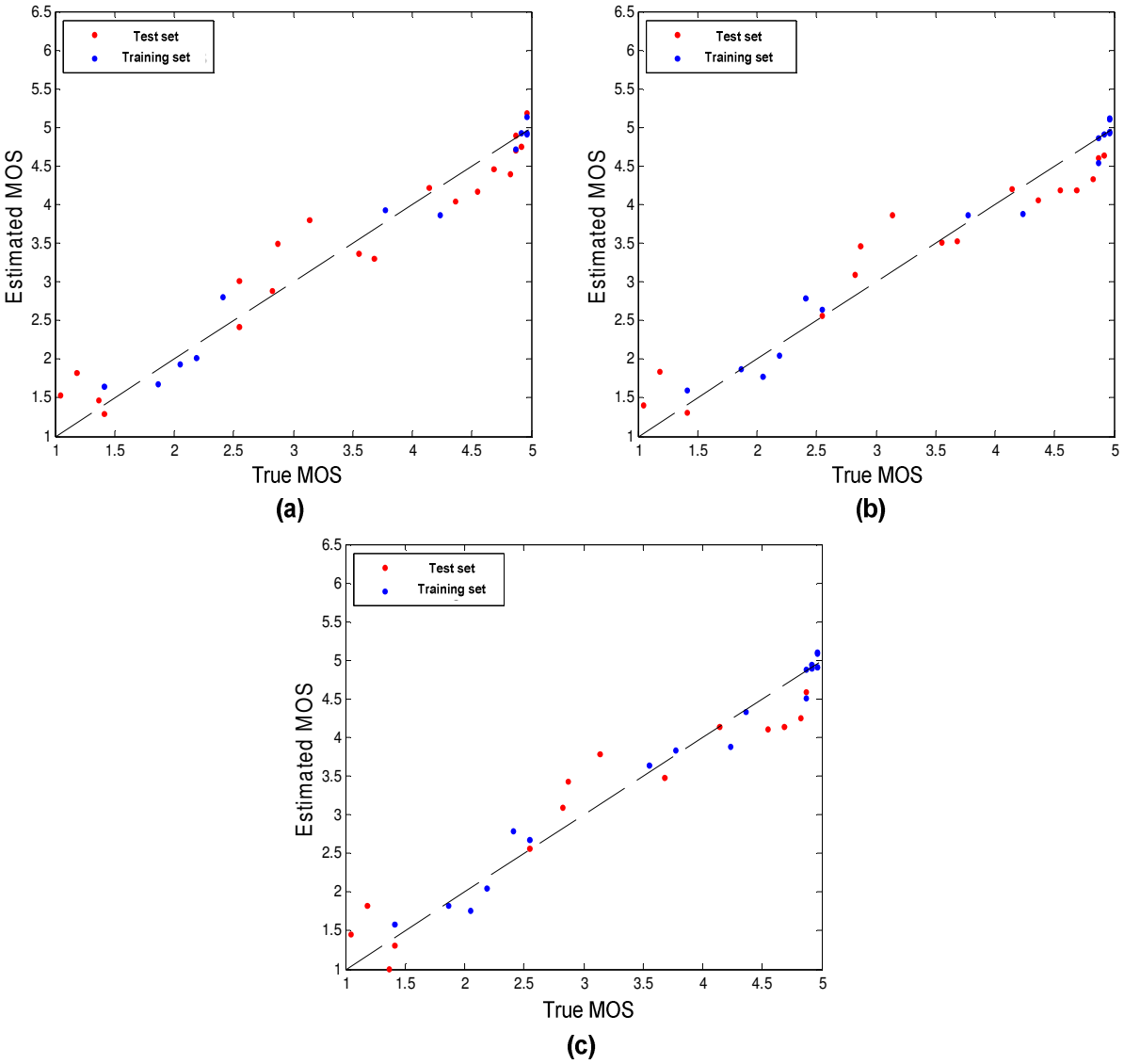


Figure 4.15: MOS estimation result for MPEG-2 using the estimated MSE: (a) 12 training/20 test video sequences; (b) 15 training/17 test video sequences; (c) 18 training/14 test video sequences

Tables 4.7.a) and b) present the results related with the RMS, the Outlier Ratio as well as the P_c and the S_c coefficients for MPEG-2, considering the "true" MSE and the estimated MSE.

Table 4.7: Model performance analysis for MPEG-2 using: (a) the "true" MSE; (b) the estimated MSE

(a)

	Training				Test				Global (Training+Test)			
Training/Test	**RMS**	**Outlier Ratio**	P_c	S_c	**RMS**	**Outlier Ratio**	P_c	S_c	**RMS**	**Outlier Ratio**	P_c	S_c
12 /20	**0.4003**	**0.0833**	**0.9570**	**0.8601**	**0.6043**	**0.1500**	**0.9023**	**0.9459**	**0.5369**	**0.1250**	**0.9221**	**0.9359**
15 /17	**0.4120**	**0.1333**	**0.9524**	**0.8857**	**0.5616**	**0.1176**	**0.9367**	**0.9314**	**0.4971**	**0.1250**	**0.9359**	**0.9545**
18 /14	**0.3998**	**0.2222**	**0.9496**	**0.9112**	**0.5602**	**0.1429**	**0.9477**	**0.9341**	**0.4767**	**0.1875**	**0.9410**	**0.9589**

(b)

	Training				Test				Global (Training+Test)			
Training/Test	**RMS**	**Outlier Ratio**	P_c	S_c	**RMS**	**Outlier Ratio**	P_c	S_c	**RMS**	**Outlier Ratio**	P_c	S_c
12 /20	**0.2068**	**0.0000**	**0.9887**	**0.9441**	**0.3546**	**0.0500**	**0.9691**	**0.9579**	**0.3076**	**0.0313**	**0.9757**	**0.9787**
15 /17	**0.1964**	**0.0000**	**0.9894**	**0.9821**	**0.3876**	**0.0000**	**0.9611**	**0.9583**	**0.3129**	**0.0000**	**0.9740**	**0.9839**
18 /14	**0.1815**	**0.0000**	**0.9898**	**0.9752**	**0.4205**	**0.0000**	**0.9561**	**0.9560**	**0.3097**	**0.0000**	**0.9745**	**0.9809**

According to Tables 4.7.a) and b), similarly with the high complexity model based on H.264 compressed video sequences, the metrics performance results are similar. However, based on the values taken by RMS, Outlier Ratio, P_c and S_c, it is possible to perceive a noticeable improvement when the MOS prediction model uses the estimated MSE instead of the "true" MSE computed using the original and the degraded videos.

Considering these results, it is possible to conclude that independently on how the MSE was computed, the model performance results allow to validate it.

4.4.3 Features space reduction with PCA

In order to reduce the model dimensionality without sacrificing the model accuracy, the method based on PCA was applied. As it was described in section 4.2.5, this method has the goal of reducing the number of features used to estimate the MOS without losing the main information and consequently without losing the model's accuracy.

The main difference between this section and the previous one is the fact that in addition to the high complexity model (4.6) described in sub-section 4.2.4, the PCA is applied in order to reduce the correlation between the features used to estimate the MOS. It is important to remark that when the PCA is applied, there should be a compromise between the number of reduced features and the model performance results. Thus, there is an ideal number of reduced features for which the performance is maximized, and as consequence, the number of features will be lower than the ones used in the original model (model before applying the PCA), as well as the model performance results can be similar to those presented by the original model. At the end, after the correlation between

features has been reduced, the number of features used to estimate the MOS, corresponds to the ones that have more relevance to the prediction.

After the application of the PCA to the group of features described in 4.2.2, the same strategy taken in previous section to compute the regression weights was followed. Since in many practical and real life video service applications, the original video sequences are not accessible at the user end side, in this sub-section only the estimated MSE between the reference and degraded video sequence was considered on the MOS prediction model presented in (4.6).

Table 4.8 exhibits the results of the regression weights for both compression standards, H.264 and MPEG-2.

Table 4.8: Regression weights for the high complexity using the estimated MSE model after applying PCA: (a) for H.264 and (b) MPEG-2

(a)

Number of features; Training/Test	β_0	β_1	β_2	β_3	β_4	β_5
4 features 12/20	3.8939	0.7152	-0.2102	-0.2109	0.2424	-
5 features 15 /17	3.5303	-0.0540	-0.8739	-0.2926	-0.1567	0.0291
5 features 18 /14	3.4722	0.0004	-0.9011	-0.2191	-0.1439	-0.0273

(b)

Number of features; Training/Test	β_0	β_1	β_2	β_3	β_4	β_5	β_6
5 features 12 /20	3.5455	0.1862	-0.8497	-0.1032	-0.0374	-0.2412	-
6 features 15 /17	3.6606	0.4569	-0.7339	-0.0364	0.0070	-0.1129	-0.0412
4 features 18 /14	3.7626	0.6226	-0.5308	-0.0566	0.0202	-	-

From Tables 4.8.a) and b), it is possible to verify that, contrarily to what was observed in the results of sections 4.4.1 and 4.4.2, the regression parameters present homogeneous values. However, there is one regression weight that stands out from both tables, which is β_2.

After having computed the regression weights, the MOS was estimated using a regression model similar to (4.6) but adapted to the number of features after applying the PCA.

Figures 4.16 and 4.17 display the results of estimated MOS vs “true” MOS for H.264 and MPEG-2, taking into account only the main features, *i.e.,* features that resulted from the selection made by the PCA.

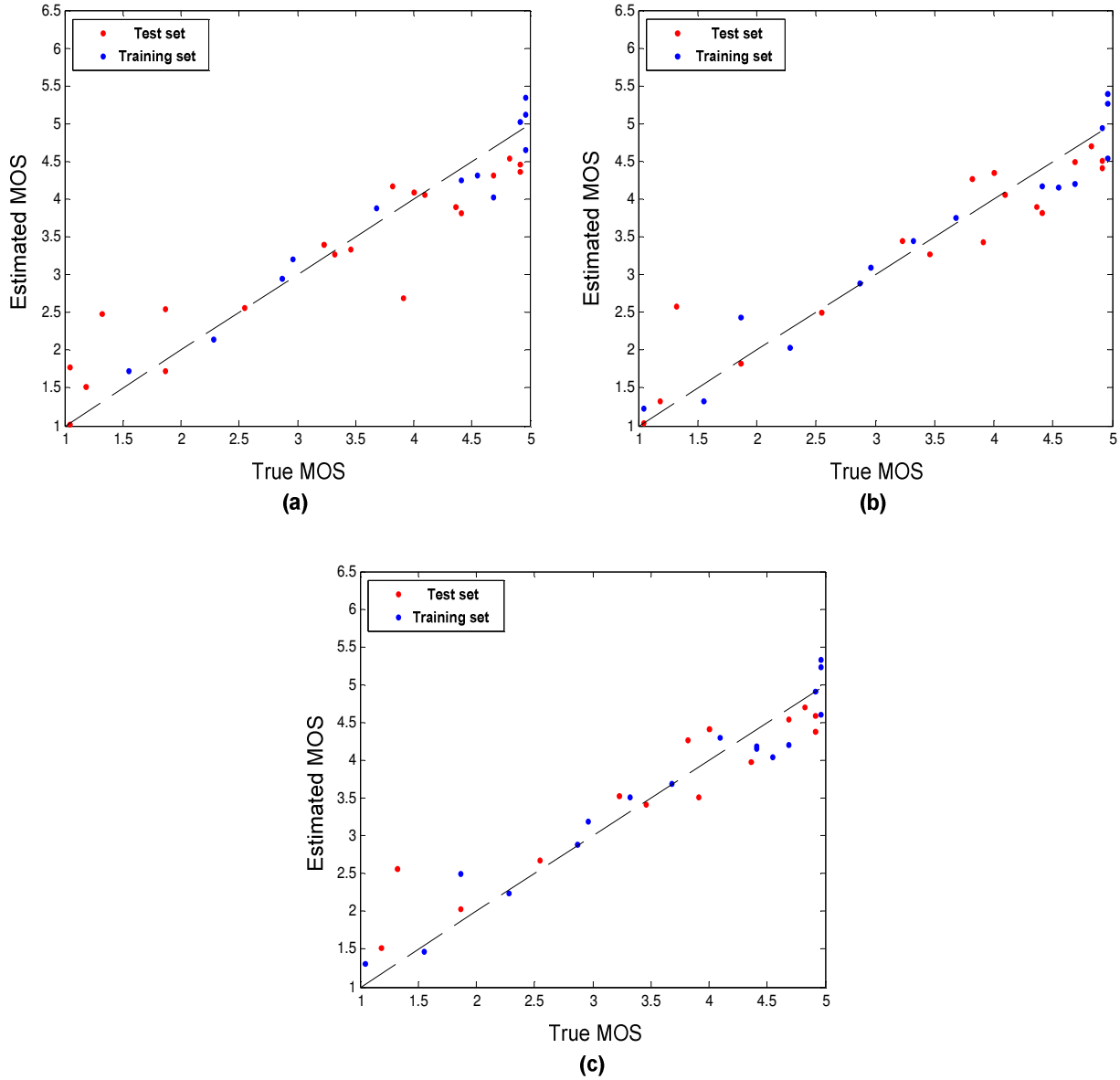


Figure 4.16: MOS estimation result for H.264 using the estimated MSE after using the PCA method: (a) 12 training/20 test video sequences; (b) 15 training/17 test video sequences; (c) 18 training/14 test video sequences

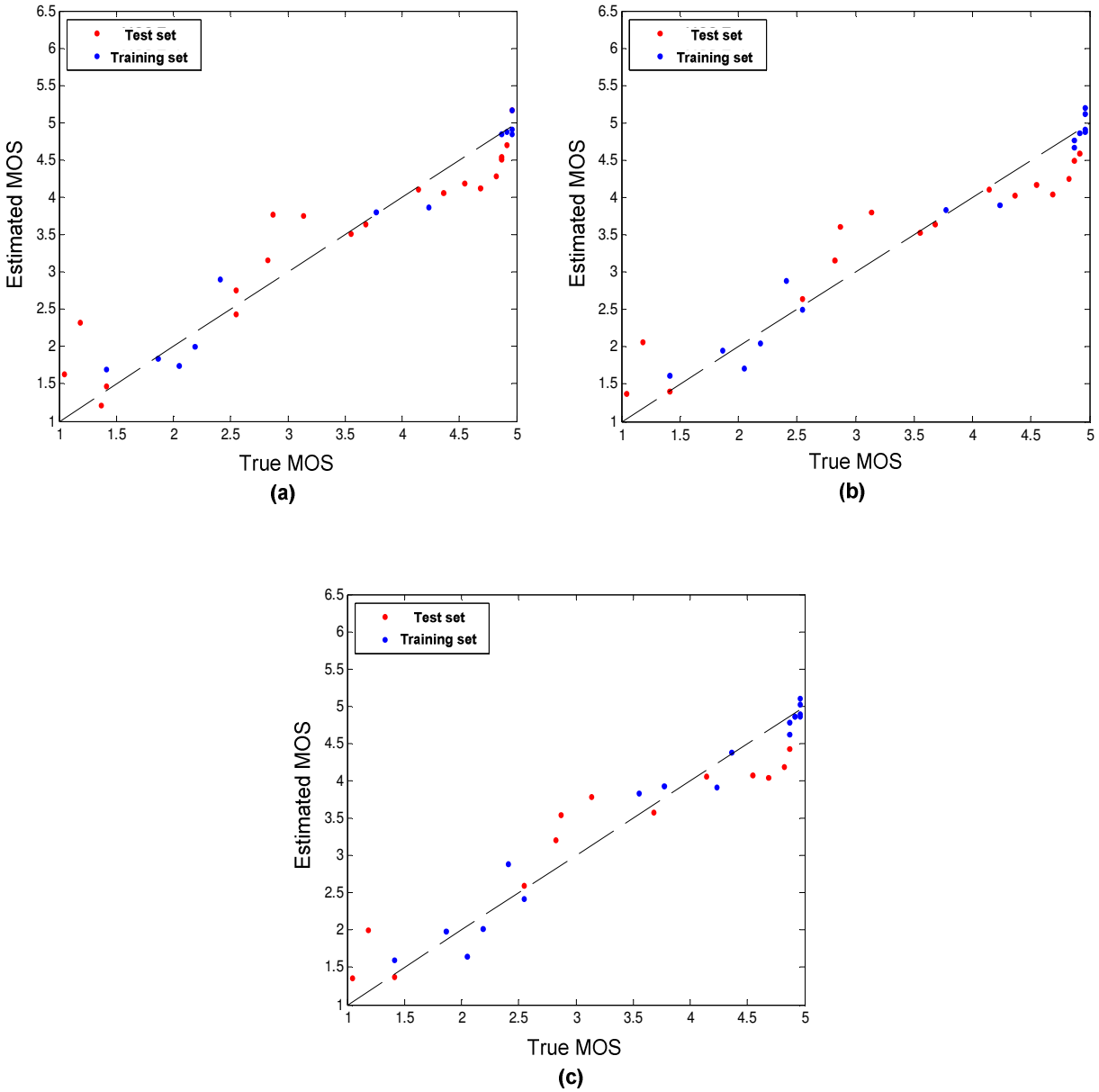


Figure 4.17: MOS estimation result for MPEG-2 using the estimated MSE after applying the PCA method: (a) 12 training/20 test video sequences; (b) 15 training/17 test video sequences; (c) 18 training/14 test video sequences

Based on Figures 4.16 and 4.17, it is possible to verify, that comparatively to Figures 4.13 and 4.15, the model accuracy seems not to be affected by the PCA. In fact, in some cases, after reducing the features redundancy, the $\hat{MOS}$ seems to be closer to the "true" MOS than the $\hat{MOS}$ computed without using the PCA.

Nevertheless, it is necessary to perform a more rigorous analysis than the one performed so far, in order to carry out a more reliable comparison between the results provided by the two methods, *i.e.*, the results achieved before and after applying the PCA method.

Thus, in order to properly evaluate the model performance, Tables 4.9.a) and b) present the results related with the RMS, the Outlier Ratio as well as the P_c and the S_c coefficients for H.264 and MPEG-2 after applying the PCA method.

Table 4.9: Metrics performance after applying the PCA for: (a) H.264 and (b) MPEG-2

(a)

	Training				Test				Global (Training+Test)			
Number of features; Training/Test	**RMS**	**Outlier Ratio**	P_c	S_c	**RMS**	**Outlier Ratio**	P_c	S_c	**RMS**	**Outlier Ratio**	P_c	S_c
4 features 12/20	0.2820	0.0000	0.9694	0.9371	0.5211	0.1500	0.9349	0.9233	0.4467	0.0938	0.9450	0.9509
5 features 15 /17	0.3078	0.0667	0.9728	0.9714	0.4386	0.0588	0.9471	0.9142	0.3829	0.0625	0.9581	0.9542
5 features 18 /14	0.3080	0.0000	0.9743	0.9567	0.4553	0.0714	0.9484	0.9165	0.3796	0.0313	0.9597	0.9597

(b)

	Training				Test				Global (Training+Test)			
Number of features; Training/Test	**RMS**	**Outlier Ratio**	P_c	S_c	**RMS**	**Outlier Ratio**	P_c	S_c	**RMS**	**Outlier Ratio**	P_c	S_c
5 features 12/20	0.2286	0.0000	0.9862	0.9650	0.4556	0.1000	0.9502	0.9699	0.3864	0.0625	0.9618	0.9765
6 features 15 /17	0.2113	0.0000	0.9877	0.9893	0.4579	0.0588	0.9443	0.9632	0.3638	0.0313	0.9645	0.9806
4 features 18 /14	0.2096	0.0000	0.9864	0.9835	0.4954	0.0714	0.9349	0.9604	0.3634	0.0313	0.9641	0.9828

The first conclusion that can be drawn from Tables 4.9.a) and b), is that the metrics performance results for both compression standards, H.264 and MPEG-2, are similar. However, it is possible to notice a slight improvement when the MOS prediction model uses video sequences compressed with the MPEG-2 compression standard instead of H.264 compressed video sequences.

According to Tables 4.9.a) and 4.5.b) for H.264 as well as Tables 4.9.b) and 4.7.b) for MPEG-2, the model performance results, are identical before and after applying the PCA. This result it is expectable, since the goal of PCA is to reduce the number of features, removing the features that show redundancy with each other, and without losing the main information as well as in the model accuracy.

4.4.4 Comparison with related work

This thesis presents two No Reference (NR) video quality estimation models: the low complexity model, based on a small set of video features, and a high complexity model based, on the same set of features with the inclusion of an additional metric, the MSE. Traditionally, a model that incorporates the MSE, *e.g.*, the PSNR, is not considered a NR model since the original video is needed to compute this metric. However, in this project, the MSE is estimated through the model developed by Brandão [BQ08b]. Although there is an increase on the system complexity, it is interesting to evaluate the influence of this feature on the accuracy of the MOS estimation. Similarly approaches, namely RR and NR models, in the sense that simple video features like blocking and blurring are also used, were proposed in [OD07] and in [KOD09], respectively, achieving in both cases slight inferior results for all VQEG measurements.

Chapter 5

Conclusions and Future Directions

The aim of this thesis was to develop an objective metric capable of predicting the MOS of compressed video sequences based only on NR features, *i.e.*, features available at the receiver side.

In order to develop a MOS prediction model approaching the behaviour of human visual system in video quality evaluation, subjective tests data were required to calibrate and validate the model. Since the majority of subjective results (e.g. those produced in MPEG groups) are only available for a restrict group of persons, this thesis built its own database. In fact, the production of this database of video sequences and associated MOS, wins a new dimension of importance since the subjective results as well as of all type of information related with them, can be used in future works by people who has interest in the video quality evaluation field.

In what concerns the objective quality evaluation, two new objective video quality assessment metrics were proposed. These models combine a small set of features extracted from video sequences available at the user side, in order to predict the MOS given by the observers during the subjective tests. The first considered approach - the low complexity model - was based on simple video features like the bitrate, the global spatial and temporal activities and the spatial and temporal activities variance, computed from the received video. The second approach - the high complexity model - also includes the MSE metric, which is estimated without the need of the reference video

[BQ08b], in order to maintain the NR property. Although the inclusion of the MSE as a feature increases the system computational complexity, it was of interest to evaluate its influence in the accuracy of the MOS estimation.

The subjective tests and the objective quality evaluation were conducted using two different compression standards, the MPEG-2 and the H.264/AVC.

The models' ability to predict subjective assessment of video quality was quantitatively evaluated using three measures: the prediction accuracy (Pearson coefficient), the prediction monotonicity (Spearman coefficient) and the prediction consistency (Outlier Ratio coefficient). Furthermore, the RMS was computed in order to provide a better perception of the MOS error estimation.

Based on the model performance results presented in chapter 4, it is possible to conclude that the two approaches are capable to correctly modelling the human visual system in video quality evaluation. However, the high complexity model shows to be the more accurate, due to the inclusion of the MSE feature. Although this second approach has the downturn of a higher computational complexity than the first one, this strategy is justified since it will improve the model accuracy and consequently approach the MOS prediction model precision.

In order to simplify the MOS prediction model by reducing the number of features used to estimate the MOS and, as a consequence, by removing the redundancy among features, a method based on PCA was conducted. It was verified that the model accuracy is not affected, although in this case the number of features to estimate the MOS were inferior to the number of features used by the original model.

For future work and in order to enhance the MOS prediction model proposed in this thesis a spatio-temporal model of the human visual system could be explicitly taken into account (for instance, as an explicit weigh of the MSE measure). Another possible enhancement could be made in the regression model chosen to estimate the MOS, if better adjusted to each feature.

References

[Bist05] Bistrom J., “Comparing video codec evaluation methods for handheld digital TV”, in T-111.590 Research Seminar on Digital Media, 2005.

[BQ08a] Brandão T. and Queluz P., "No-reference image quality assessment based on DCT domain statistics", Signal Processing, Vol. 88, No. 4, pp. 822 - 833, April, 2008.

[BQ08b] Brandão T. and Queluz P., “No-reference PSNR estimation algorithm for H.264 encoded video sequences,” in Proc. of EUSIPCO - European Signal Processing Conference, Lausanne, Switzerland, August 2008.

[GGC01] Grgic S., Grgic M., Cihlar B., “Performance analysis of image compression using wavelets”, IEEE Transactions on Industrial Electronics, Vol 48, No 3, 2001.

[HBL02] Hekstra A., Beerends J., Ledermann D., “PVQM – A perceptual video quality measure – Signal Processing: Image Communication”, 17(10):781–798, 2002.

[ITU98] ITU-R BT. 500-9, “Methodology for the subjective assessment of the quality of television pictures”, 1998.

[ITU99] ITU-T P.910, “Subjective video quality assessment methods for multimedia applications”, 1999.

[ITU04] ITU-T, “Objective perceptual assessment of video quality: full reference television”, 2004.

[ITU08a] ITU-T J.246, “Perceptual visual quality measurement techniques for multimedia services over digital cable television networks in the presence of a reduced bandwidth reference”, 2008.

[ITU08b] ITU-T J.247, “Objective perceptual multimedia video quality measurement in the presence of a full reference”, 2008.

[OD07] Oelbaum T., Diepold K., “A reduced reference video quality metric for AVC/H.264”, in proc. of EUSIPCO – European Signal Processing Conference, Poznan, Poland, September 2007, pp. 1265–1269.

[KOD09] Keimel C., Oelbaum T., Diepold K., "No-reference video quality evaluation for high-definition video", icassp, pp.1145-1148, 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, 2009.

[Mira02] Miras D., “A survey on network QoS needs of advanced internet applications”, Technical report, Internet2 - QoS Working. Group, 2002.

[Pear99] Pearson D., ”Viewer response to time-varying video quality”, in Proceedings of the SPIE Human Vision and Electronic Imaging, 3299, pp. 16–25, (San Jose, CA), January 1999.

[RNR07] Ries, M., Nemethova, O. and Rupp, M., ”Motion based reference-free quality estimation for H.264/AVC video streaming”, in Proc. of IEEE Int. Symp. on Wireless Pervasive Computing (ISWPC), San Juan, Puerto Rico, US, Feb. 2007.

[Wats93] Watson A., “DCT quantization matrices optimized for individual images,” in Proc. of SPIE Human Vision, Visual Processing, and Digital Display IV, S. Jose, USA, 1993.

[WBSS04] Wang Z., Bovik A., Sheikh H. and Simoncelli E., “Image Quality Assessment: From Error Visibility to Structural Similarity”, IEEE T. Image Processing, Vol. 13, No. 4, pp. 600-612, April 2004.

[Wink07] Winkler S., “Video quality and beyond,” in proc. of EUSIPCO - European Signal Processing Conference, Poznan, Poland, September 2007.

[WP99] Wolf S. and Pinson M., "Spatial-temporal distortion metrics for in-service quality monitoring of any digital video system", in Proceedings of the Multimedia Systems and Applications II, vol. 3845 of Proceedings of SPIE, pp. 266-277, Boston, Mass, USA, September 1999.

[WP07] Wolf S. and Pinson M., “Application of the NTIA general video quality metric (VQM) to HDTV quality monitoring”, in Proceedings of The Third International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM), Scottsdale, AZ, USA, January 2007.

[WR06] Wu H., Rao K., “Digital video image quality and perceptual coding”, CRC Press, Taylor & Francis Group, Florida, USA, 2006.

[WSB03] Wang Z., Sheikh H., Bovik A., “Objective video quality assessment, in The Handbook of Video Databases: Design and Applications”, B. Furht and O. Marqure, Editors. 2003, CRC Press. p.1041-1078.