FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Joint Source Video Coding

Joint Rate Control for H.264/AVC Video Coding

A dissertation submitted in partial satisfaction of the

requirements for the degree

Doctor of Philosophy

in

Electrical and Computer Engineering

by

Luís Miguel Lopes Teixeira

Jury

President: Doutor Artur Pimenta Alves

Member: Doutor António José Nunes Navarro Rodrigues

Member: Doutor Aurélio Joaquim de Castro Campilho

Member: Doutor Fernando Manuel Bernardo Pereira

Member: Doutor Luís António Pereira de Meneses Corte-Real

2012

Acknowledgements

Although writing most of the time is a solitary activity, in providing the content, the ideas, the motivation and the strength to finalize any text, one relies on many others. Writing this section of the thesis gives me a formal opportunity to thank people who have supported me and consequently, had an influence on the accomplishment of this work. I am sincerely indebted and thankful to all of them.

First, my deepest gratitude is due to Luís Corte-Real, for his permanent guidance, productive feedback and moral support the whole time. Luís willingly provided additional advice and constructive comments as well confidence and support at times where things were not looking rosy. To Artur Pimenta Alves and José Ruela, who gave me the opportunity to work in such a unique research centre as INESC PORTO, from which I have learnt a lot. I am grateful to all my friends from INESC PORTO, for numerous inspiring discussions, help with my research and general advice.

I would also like to express my gratitude to the Fundação para a Ciência e Tecnologia (FCT), for the financial support given through the PRAXIS XXI scholarship program.

Lastly, I would like to thank to my friends, especially Ana, Carlos Caires, Carlos Ruiz and José, for their company and support. Finally, I am eternally grateful to my mother, brothers and Joana for their understanding, warmth, infinite patience and encouragement.

Abstract

In video broadcast systems, the principal objective is to obtain the best possible visual quality at a specific rate constraint and channel settings. Traditional methods for broadcasting multiple digital video bitstreams over a single-channel use a straightforward approach: they divide the existing capacity of a broadcasting channel equally among all programmes, and then they independently encode each video programme at a constant bit rate (CBR). In this dissertation, we focus on how to design efficient rate control schemes for the joint coding of video bitstreams using H.264/AVC video encoders in such a way that the sum of all the bit rates meets the negotiated connection parameters. A possible solution is to encode each video bitstream using VBR video encoding algorithms, guaranteeing that the permissible range of variation of the bit rate of each source is restricted by at least the capacity of the channel. In order to estimate the impact in terms of video quality of encoded video bitstreams, several issues have been studied: the criteria to allocate bandwidth between the different video programmes and how visual quality is assessed.

The way video quality is measured is vital for designing video broadcast systems that potentially result in the degradation of the visual quality. Such metrics have the potential to allow designers to optimise video broadcast systems to deliver higher quality while decreasing system costs. For decades, the mean squared error (MSE) has been the dominant objective video quality metric. Nevertheless, on many occasions the MSE exhibits weak performance and has been extensively criticized for serious shortcomings, mainly when dealing with perceptually significant signals. Research on how perceptual video quality metrics, metrics that incorporate human visual perception, can be used in the coding process is rather important. In the present dissertation, two video quality metrics that use the structural properties of human vision were selected, the Structural SIMilarity (SSIM) index and the JND (Just Noticeable Distortion). A novel approach to a joint video source coding, based on perceptual distortion, is thus proposed.

The H.264/AVC standard specifies the decoding process and the bit-stream syntaxes accordingly allowing research towards the optimization of the encoding process concerning coding performance improvement and complexity reduction. Rate-distortion (R-D) based methods are frequently used to enhance and smooth video quality. However the lack of R-D models for joint video encoders, using perceptual video quality metrics, confines their application to joint video coding. To provide a more accurate estimation, the rate distortion

relation based on the characteristics of the H.264/AVC coding to support the best video quality perception was researched and modeled. One important outcome is that the models deliver good performance in terms of quality prediction accuracy and strong correlation with subjective ratings. So, one can exploit the accuracy of quality prediction of existing perceptual video quality metrics to facilitate joint coding decisions. These results demonstrate that subjective metrics can be incorporated in joint source coding systems resulting in a perceived quality improvement.

Experiments have been conducted to validate the performance of the proposed algorithms. Objective and subjective video quality results show that the proposed algorithms outperform current state-of-the-art methods in the simulations. With suitable bit rate allocation, the fluctuation of objective video quality is reduced in a joint video coding process.

In summary, the main contributions of this thesis are as follows: modelling of R-D of H.264/AVC using perceptual means to assess the distortion; development of joint source coding algorithms for controlling a statistical multiplexing process of different streams into a fixed bandwidth channel that incorporate perceptual information; study how joint coding results correlate with subjective quality assessment.

Resumo

Os sistemas de transmissão de vídeo comprimido tem como objetivo principal difundir um ou mais fluxos de vídeo digital codificados com a melhor qualidade visual possível, de acordo com as definições do canal de transmissão e requisitos da taxa de transmissão de bits (débito binário). Os métodos tradicionais de transmissão de um ou mais fluxos de vídeo comprimido num único canal utilizam uma abordagem direta: a capacidade do canal de transmissão é dividida de forma equitativa pelos diferentes fluxos de vídeo digital e posteriormente cada fluxo de vídeo digital é codificado a débito binário constante.

Esta dissertação tem como objetivo o estudo de sistemas de controle de débito binário quando vários fluxos de vídeo digital são codificados simultaneamente e transmitidos num único canal, usando codificadores H.264/AVC. O valor da soma do débito binário de cada fluxo de vídeo comprimido deve respeitar os valores dos parâmetros da ligação negociada. Para a resolução deste problema é proposto e especificado um sistema que permite codificar vários fluxos de vídeo digital, usando algoritmos de codificação de débito binário variável. O sistema proposto limita a variação do débito binário de cada fluxo de vídeo comprimido de acordo com a capacidade do canal. É apresentado o desempenho da solução proposta para diferentes critérios de distribuição da largura de banda entre os diferentes fluxos de vídeo comprimidos. A avaliação da qualidade visual final é avaliada recorrendo a métodos de avaliação objetivos e subjetivos.

A escolha do procedimento para avaliação da qualidade visual é determinante para o desenho de sistemas de transmissão de vídeo digital. Esta escolha condiciona o desenvolvimento de sistemas capazes de simultaneamente otimizar a qualidade visual e garantir a redução da complexidade do sistema. Durante décadas, o erro médio quadrado (MSE) foi a métrica de qualidade vídeo dominante. No entanto, em muitas situações, o MSE exibe um desempenho fraco e tem sido extensivamente criticado por falhas graves, principalmente quando se lida com sinais perceptualmente relevantes. Desta forma é relevante estudar métodos de incorporar as métricas de qualidade percetual de vídeo no processo de codificação, métricas estas que integram a perceção visual humana no processo de aferição da qualidade visual. Para este estudo, foram selecionadas duas métricas de qualidade de vídeo que utilizam as propriedades estruturais da visão humana, a similaridade estrutural (SSIM) e o índice de JND (Just

Noticeable Distortion). A proposta final consiste no desenvolvimento de métodos de codificação conjunta de fluxos de bits de vídeo, baseada na distorção percetiva.

A norma H.264/AVC especifica somente o processo de descodificação, a sintaxe e semântica do fluxo de bits codificado, incidindo a investigação na otimização do processo de codificação relativamente ao desempenho e à redução da complexidade associada. Os métodos de Rate-Distortion (RD) são frequentemente utilizados para melhorar e uniformizar a qualidade visual. No entanto, a ausência de modelos Rate-Distortion em cenários de codificação conjunta de fontes de vídeo, utilizando métricas de qualidade percetual de vídeo, limita o seu uso. Para ultrapassar esta limitação, procedeu-se à análise e modelação da relação entre o débito binário e a distorção associada a métricas de qualidade percetual usando fluxos de vídeo com base nas características de um codificador H.264/AVC. Os modelos Rate-Distortion obtidos apresentam um elevado desempenho em termos de precisão da predição da qualidade dos sinais de vídeo digital e uma forte correlação com processos de avaliação da qualidade usando metodologias de avaliação subjetiva. Com base nestes resultados, foram incorporadas métricas de qualidade percetual no processo de codificação conjunta de fluxos de vídeo. A análise do desempenho destes sistemas foi baseada num conjunto de simulações para cada um dos algoritmos propostos. Os resultados obtidos foram avaliados usando métricas objetivas e subjetivas de avaliação da qualidade, evidenciam um desempenho superior os atuais métodos. Uma melhor alocação do débito binário entre os diferentes fluxos de vídeo permite ainda reduzir a flutuação da qualidade de vídeo, num fluxo de vídeo e entre fluxos de vídeo.

Em síntese, as principais contribuições desta dissertação são: a modelação Rate-Distortion de um codificador H.264/AVC utilização métricas de qualidade percetual, o desenvolvimento de algoritmos de codificação conjunta, que incorporem informação percetual, para controlar o processo de multiplexagem estatística de diferentes fluxos de vídeo num canal único de difusão, e o estudo da correlação entre os resultados da codificação conjunta e a avaliação subjetiva dos resultados.

Resumée

En ce qui concerne les systèmes de diffusion vidéo, le principal objectif consiste à obtenir la meilleure qualité visuelle possible, à un taux de débit binaire spécifique et selon les définitions du canal. Les méthodes traditionnelles de diffusion de multiples trains de bits de vidéo numérique sur une seule chaîne ont recours à une approche directe : elles répartissent la capacité existante d'un canal de radiodiffusion de manière équitable entre tous les programmes, puis encodent chaque programme vidéo de manière indépendante à un débit binaire constant (CBR). Le présent exposé se concentre sur la conception de schémas efficaces de contrôle de débit pour le codage conjoint de trains de bits vidéo utilisant des encodeurs de vidéo H.264/AVC de sorte à ce que la somme de tous les débits binaires respecte les paramètres de la connexion négociés. Une solution consiste à encoder chaque train de bit vidéo au moyen d'algorithmes d'encodage vidéo au débit binaire variable (VBR), en assurant ainsi que la capacité du canal restreint l'intervalle de variation du débit binaire de chaque source. Afin d'estimer l'impact en termes de qualité vidéo des trains de bits vidéo encodés, nous avons analysé plusieurs points : les critères pour allouer la largeur de bande entre les différents programmes vidéo et la manière dont la qualité visuelle est évaluée.

La manière dont la qualité de la vidéo est mesurée est essentielle pour la conception de systèmes de diffusion vidéo pouvant mener à la dégradation de la qualité visuelle. Ces métriques permettent aux concepteurs d'optimiser les systèmes de diffusion de vidéo afin d'offrir une meilleure qualité, tout en réduisant les coûts du système. Pendant plusieurs décennies, l'erreur quadratique moyenne (EQM) a été la métrique de qualité vidéo dominante. Néanmoins, l'EQM montre, à de nombreuses reprises, une faible performance et a été largement critiquée en raison de graves lacunes, notamment lors du traitement de signaux pertinents du point de vue perceptuel. La recherche sur la manière d'utiliser des métriques de qualité perceptuelle de vidéo (métriques intégrant la perception visuelle humaine) dans le processus de codage est importante. Dans le présent exposé, nous avons eu recours à deux métriques de qualité vidéo reposant sur les propriétés structurelles de la vision humaine : l'index de similitude structurale (SSIM) et la JND (Just Noticeable Distortion). Nous proposons ainsi une nouvelle approche pour un codage conjoint de flux de bits vidéo, reposant sur la distorsion.

La norme H.264/AVC spécifie le processus de décodage et les syntaxes de trains de bits, permettant ainsi la recherche visant à optimiser le processus d'encodage quant à l'amélioration

de la performance de codage et à la réduction de la complexité. Nous avons fréquemment recours à des méthodes reposant sur le degré de distorsion (R-D) afin d'améliorer et d'uniformiser la qualité de la vidéo. Toutefois, le manque de modèles R-D pour les encodeurs vidéo conjoints, utilisant des métriques perceptuelles de qualité vidéo, limite leur application au codage vidéo conjoint. Ainsi, afin d'offrir une estimative plus précise et une meilleure perception de la qualité vidéo, nous avons analysé et modelé le rapport du degré de distorsion reposant sur les caractéristiques du codage H.264/AVC. D'après le résultat obtenu, les modèles offrent une performance correcte en termes de précision de prévision de qualité et une forte corrélation avec les classifications subjectives. Il est donc possible d'exploiter la précision de la prévision de la qualité des actuelles métriques perceptuelles de qualité de vidéo pour faciliter les décisions de codage conjoint. Ces résultats montrent que des métriques subjectives peuvent être intégrées aux systèmes de codage conjoint de sources, ce qui mène à une amélioration de la qualité perceptuelle.

Des essais ont été effectués afin de valider la performance des algorithmes proposés. Les résultats objectifs et subjectifs des simulations de la qualité vidéo montrent que les algorithmes proposés présentent un résultat supérieur à celui des méthodes actuelles. Une distribution appropriée du débit binaire permet la réduction de la fluctuation de la qualité vidéo au moyen du processus de codage vidéo conjoint.

En résumé, les principales contributions de cette thèse sont les suivantes : le modelage RD d'un encodeur H. 264/AVC en ayant recours à des moyens perceptuels afin d'évaluer la distorsion ; le développement d'algorithmes de codage conjoint de source pour le contrôle du processus de multiplexage statistique de différents trains de bits en un canal à largeur de bande fixe intégrant des informations perceptuelles ; l'analyse de la manière dont les résultats de codage conjoint ont un rapport avec l'évaluation subjective de la qualité.

He who becomes the slave of habit, who follows the same routes
every day, who never changes pace, who does not risk and
change the color of his clothes, who does not speak and
does not experience, dies slowly.

He or she who shuns passion, who prefers black on white,
dotting ones "i's" rather than a bundle of emotions,
the kind that make your eyes glimmer,
that turn a yawn into a smile, that make the heart pound
in the face of mistakes and feelings, dies slowly.

He or she who does not turn things topsy-turvy,
who is unhappy at work, who does not risk certainty
for uncertainty, to thus follow a dream, those who do not forego
sound advice at least once in their lives, die slowly.

He who does not travel, who does not read,
who does not listen to music, who does not find grace in himself,
she who does not find grace in herself, dies slowly.

He who slowly destroys his own self-esteem,
who does not allow himself to be helped,
who spends days on end complaining about his own bad luck,
about the rain that never stops, dies slowly.

He or she who abandon a project before starting it,
who fail to ask questions on subjects he doesn't know,
he or she who don't reply when they are asked something
they do know, die slowly.

Let's try and avoid death in small doses, reminding oneself
that being alive requires an effort far greater than the simple fact of breathing.

Only a burning patience will lead
to the attainment of a splendid happiness.

Pablo Neruda

List of Acronyms

The following is a description of the acronyms used in this dissertation

AAC	Advanced Audio Coding.
ACR	Absolute Category Rating.
AFX	Animation Framework eXtension.
AI	Artificial Intelligence.
ALS	Audio Lossless Coding.
ANSI	American National Standards Institute.
ATIS	Alliance for Telecommunications Industry Solutions.
AVC	Advanced Video Coding.
AVI	Audio Video Interleave.
CABAC	Context-Adaptive Binary Arithmetic Coding.
CAVLC	Context-Adaptive Variable-Length Coding.
CBR	Constant Bit Rate.
CCIR	Comité Consultatif International des Radio Communications, or International Radio Consultative Committee).
CCITT	Comité Consultatif International des Téléphonique et Télégraphique, (International Telegraph and Telephone Consultative Committee)
CD	Compact Disc.
CDF	Cumulative Distribution Function.
CFF	Compact Font Format.
CI	Confidence Interval.
CIE	Commission International de l'Eclairage.
CIF	Common Intermediate Format.
CoV	Coefficient of Variation.
CPB	Coded Picture Buffer.
CQ-VBR	Constant Quality - Variable Bit Rate.
CSF	Contrast Sensitivity Function.
DCR	Degradation Category Rating.
DCT	Discrete Cosine Transform.
DDL	Description Definition Language.
DMIF	Delivery Multimedia Integration Framework.
DMOS	Difference Mean Opinion Score.
DP	Dynamic Programming.
DPB	Decoded Picture Buffer.
DPCM	Differential Pulse Code Modulation
D-Q	Distortion-Quantisation.
DRM	Digital rights management.
DS	Description Scheme.
DSCQS	Double Stimulus Continuous Quality Scale.
DSIS	Double Stimulus Impairment Scale.
DSL	Digital Subscriber Line.
DSM	Digital Storage Media.
DVB	Digital Video Broadcasting.
DVB-C	Digital Video Broadcasting Cable.
DVB-H	Digital Video Broadcasting-Handheld.
DVB-S	Digital Video Broadcasting Satellite.
DVB-T	Digital Video Broadcasting Terrestrial.
DVD	Digital Versatile Disc.

EBU	European Broadcasting Union.
FR	Full Reference Model.
FR-TV	Full Reference Television.
GOP	Group of Pictures.
GOV	Group of VOP.
HD DVD	High Density DVD, or High-Definition DVD.
HD	High Definition video format.
HDTV	High-Definition Television.
HP	High Profile.
HRC	Hypothetical Reference Circuit.
HRD	Hypothetical Reference Decoder.
HSS	Hypothetical Stream Scheduler.
HVS	Human Visual System.
IDCT	Inverse Discrete Cosine Transform.
IDR	Instantaneous decoding Refresh.
IEC	International Electrotechnical Commission.
ILSC	Independent Labs and Selection Committee.
IMCP	Interlayer Motion Compensated Predictor.
IME	Interlayer Motion Estimator.
IPMP	Intellectual Property Management and Protection.
IPMP	Intellectual Property Management and Protection
IPTV	Internet Protocol Television
ISO	International Standards Organization.
ITU	International Telecommunications Union.
ITU-R	International Telecommunication Union – Radiocommunication Sector.
ITU-T	International Telecommunication Union – Telecommunication Standardisation Sector.
JND	Just Noticeable Difference.
JPEG	Joint Photographic Experts Group.
JRG	Joint Rapporteurs Group.
JTC	(ISO/IEC) Joint Technical Committee
JVCM	Joint Video Coding Multiplexer.
JVT	Joint Video Team.
KLT	Karhunen-Loeve transformation.
LASeR	Lightweight Application Scene Representation.
LBC	Low bit rate Coding.
LNG	Lateral Geniculate Nucleus.
LNP	Linear with Nonpolynomial model.
MAD	Mean of Absolute Difference.
MAD	Minimizing Average Distortion.
MB	Macroblock.
MDV	Minimizing Distortion Variation.
MMQA	Multimedia Quality Assessment.
MOS	Mean Opinion Score.
MP	Main Profile.
MPEG	Motion Picture Experts Group.
mquant	macroblock quantisation.
MSE	Mean Squared Error.
MSENL	Mean Square Error after Non-linearity.
MVO	Multiple Video Object.
MVP	Multiview Profile.
NAL	Network Abstraction Layer.
NR	No Reference Model.
NTIA	National Telecommunications and Information Administration.
NTSC	National Television System Committee.
OFF	Open Font Format.
OL-VBR	Open-Loop Variable Bit Rate.
ORD	Operational R-D.
PAL	Phase Alternate Line.

PC	Pair Comparison.
PDA	Personal Digital Assistant.
PDF	Probability Density Function.
PPD	MPEG-4 Proposal Package Description.
PPS	Picture Parameter Set.
PQR	Picture Quality Rating.
PSNR	Peak Signal to Noise Ratio.
PSPNR	Peak Signal-to-Perceptible-Noise Ratio.
PSTN	Public Switched Telephone Network.
PVS	Processed Video Sequence.
QART	Quality Recognition Tasks.
QCIF	Quarter Common Intermediate Format.
QoE	Quality of Experience
QoP	Quality of Perception
QP	Quantisation Parameter.
RDO	Rate-distortion optimisation.
RGB	Red-Green-Blue.
RMSE	Root Mean Squared Error.
R-Q	Rate-Quantisation.
RR	Reduced Reference Model.
RRNR-TV	Reduced Reference and No Reference Television.
SA	Structured Audio
SAD	Sum of Absolute Differences.
SAF	Simple Aggregation Format.
SAMVIQ	Subjective Assessment Methodology for Video Quality.
SATD	Sum of Absolute Transform Differences.
SDSCE	Simultaneous double stimulus for continuous evaluation
SDT	Signal-Detection Theory.
SDTV	Standard Definition television.
SI	Spatial Information.
SIF	Standard Intermediate Format.
SLS	Scalable Lossless Coding
SMPTE	Society of Motion Picture and Television Engineers.
SMR	Symbolic Music Representation.
SNR	Signal to Noise Ratio
SP	Simple Profile.
SPS	Sequence Parameter Set.
SRC	Scalable Rate Control.
SRC	Source Reference Channel or Circuit.
SSCQE	Single Stimulus Continuous Quality Evaluation.
SSD	Sum of Squared Differences.
SSIM	Structural Similarity Index.
SSM	Single Stimulus Methods.
STA	Spatio-Temporal Analyser.
Std	Standard Deviation.
SVC	Scalable Video Coding.
TI	Temporal Information.
TM5	MPEG-2 Test Model Version 5.
TMN10	H.263 Test Model Near-term Version 10.
TMN8	H.263 Test Model Near-term Version 8.
TR	Technical Report.
TTSI	Text-To-Speech Interface
TV	Television.
VBR	Variable Bit Rate
VBV	Video Buffer Verifier.
VC-1	Video Compression (Coding) 1.
VCEG	Video Coding Experts Group.
VCL	Video Coding Layer.
VCV	Video Complexity Verifier.

VD	Variability-Distortion.
VLC	Variable Length Code.
VLSI	Very-large-scale integration.
VM8	MPEG-4 Verification Model Version 8.
VO	Video Object.
VoIP	Voice-over-IP.
VOL	Video Object Layer.
VOP	Video Object Plane.
VQEG	Video Quality Experts Group.
VQM	Video Quality Measure.
WCWSSIM	Complex wavelet SSIM.
WT	Weighter.
YCbCr	Luminance; Chroma: Blue; Chroma: Red

Contents

List of Figures

List of Tables

Chapter 1. Introduction

Over the past twenty years, video signals have been the subject of considerable research. The advent of digital video technology has made it possible to process, broadcast, and store video streams. With the growing availability of digital transmission links, a wide range of emerging applications, such as digital TV/HDTV broadcasting, digital cinema, video conferencing or surveillance, have been developed. Currently, 48 hours of video are uploaded every minute onto YouTube databases, and over three billion videos are watched every-day on YouTube [1].

With the growing commercial interest in these products and services, the need for international audiovisual standards has emerged. The standardisation process facilitates equipment interoperability from different manufacturers. When video is being broadcasted, its quality depends on the video encoding process and the allocated bandwidth. The encoding process is fundamental since it has a huge impact on the rate-distortion performance and also on the utilization of different resources such as processing power, transmission bandwidth, and end-to-end delay of streaming service. The difficulty is even greater when the global performance over several streaming services is considered.

Digital TV, one of the most popular digital video applications, is based on the success of MPEG-2 and the Digital Video Broadcasting (DVB) standard family (Digital Video Broadcasting Terrestrial – DVB-T, Digital Video Broadcasting Satellite – DVB-S, and Digital Video Broadcasting Cable – DVB-C standard) ([2],[3],[4],[5]). Following the success of MPEG-2, the ITU-T Video Coding Experts Group (VCEG) and the ISO/IEC Moving Picture Experts Group (MPEG) have jointly developed H.264/AVC, also known as Advanced Video Coding (AVC) ([6]). H.264/AVC has achieved considerable progress regarding coding efficiency, substantially enhanced error robustness, increased flexibility and scope of applicability relative to its predecessors ([7],[8]). It covers video applications ranging from mobile services and videoconferencing to IPTV, HDTV, and HD video storage ([8],[9]). Typical target applications include broadcasts over cable, satellite, cable modem, x (of any type) digital subscriber line (xDSL), and terrestrial channels, interactive or serial storage on optical and magnetic devices such as DVD, storage and distribution of professional film and video material for content contribution and distribution, studio editing, and processing and wireless

networks ([8]). The H.264/AVC standard ([6]) has been considered ‘highly recommended’ as the video coding format for DVB-H ([10]). It is generally accepted that H.264/AVC will have a higher impact in digital video systems (such as storage, broadcast and streaming), but there are still difficulties.

According to [11] H.264/AVC compression efficiency tends to be insufficient for the wireless and wired transmission of the next-generation resolutions. As a result MPEG and VCEG have formed a Joint Collaborative Team on Video Coding (JCT-VC) to develop a successor to H.264/AVC. This standard is being referred to as High Efficiency Video Coding (HEVC) ([12], [13]). HEVC is focused on coding progressively scanned rectangular frames, and can scale from 320 x 240 pixels up to 7680 x 4320 resolution. JCT-VC integrated features from some of the best-performing HEVC proposals ([14]). First results show a reduction in bit rate requirements by half with similar subjective perceptual quality when compared with H.264/AVC, at the expense of increased computational complexity ([13]). The time schedule of JCT-VC is to publish draft versions of HEVC in 2012 and the first version of the final draft standard in January 2013. Given that HEVC is still an ongoing process with many variable elements and open questions, this thesis focuses on the existing established standards, mainly in H.264/AVC.

The video coding standards’ specifications provide only the bitstream syntax and allow flexible implementations of the encoding process. A video coding algorithm specifies how to combine standard tools to achieve efficient compression while maintaining video quality as high as possible. The combined selection of the video coding algorithm and encoding parameters has a strong impact on the quality and bit rate of the encoded video source.

The core of this dissertation is on optimizing the video coding process to enhance coding performance when various video sequences are simultaneously encoded. In order to estimate impact in terms of image quality of jointly encoded video programmes, several issues need to be studied: the criteria to allocate bandwidth between the different video programmes and the used image quality metric. Perceptual quality metrics are used in this study. Several mathematic models are investigated in order to provide support for the joint coding of video sequences in such a way that the sum of all the bit rates meets the negotiated connection parameters. In particularly, we use rate-distortion (RD) models to estimate the relationship generated by the encoding process between bits needed to encode frames and the perceived image quality. These models can be obtained analytically or empirically. In analytical modeling, the RD model is obtained by combining the statistics of the source video signal with the properties of the video encoder. In empirical modeling, the RD model is generated by an interpolation process between a set of RD points, in an approximation of the RD curve. The RD model is then used in the control of the bandwidth allocated per video sequence. After the encode process ends, the model

is updated. Using valid models enables us to estimate these parameters. The prime theme underlying all of our work is the use of statistical modeling techniques to optimize the video coding process and the use of perceptual quality metrics.

1.1 Video Coding and Rate Control

The MPEG standards family specifies the decoding process and the bit-stream syntaxes allowing research towards the optimization of the encoding process regarding coding performance improvement and complexity reduction. The objective of a video encoder is to generate the optimum perceptual video quality, or to minimise distortion, under a certain set of requirements such as channel bandwidth or storage limitations. In general, for a specific bit budget, the video encoder should optimally determine a set of the best quantisation parameters by minimizing the value of the distortion D. It is known that the quantisation parameter plays a key role in the generation of bits and distortion coding. If a video sequence is encoded using all the different quantisation parameters, then rate and quantisation error can be obtained, and it is possible to plot the rate-quantisation (R-Q) or distortion-quantisation (D-Q) curves. R-Q and *D-Q* functions characterise the rate-distortion (*R-D*) behaviour of video encoding that allows the improvement of bit allocation.

There are two main approaches to solving the optimal bit allocation problem: Lagrange's optimization ([15],[16]) and dynamic programming (DP) ([17]). However, the computational complexity of these methods is very high due to the need to determine R-D characteristics of current and future video frames. Accordingly, to obtain an estimation of the bit rate without having to implement the whole encoding process, mathematical models can estimate the bit rate or the quantisation error. Many *R-Q* and *D-Q* functions have been reported in previous studies ([18],[19],[20],[21],[22],[23],[24],[25],[26],[27],[28]). Some of these schemes were adopted in standard-compliant video coders, such as TM-5 ([19]), the test model for MPEG-2, TMN-8 ([20]), the test model for H.263, or VM-8 ([28]), the verification model for MPEG-4. The algorithms proposed in this dissertation will be implemented using the H.264/AVC standard.

To provide an accurate bandwidth estimation for the different video programmes, the rate distortion relationship will be further investigated. In particular, it will be research how perceptual-based quality metrics for image and video can be incorporated in the joint coding of multiple video streams. The study will focus on two perceptual quality image metrics, the Structural SIMilarity (SSIM) index and the JND (Just Noticeable Distortion). With suitable bit rate allocation, the fluctuation in video quality can be reduced.

1.2 Problem Statement and Objectives

Digital techniques are being used in a wide range of video applications, mainly in residential digital video services such as digital TV, Video on Demand, Internet video, etc. - each service having several programmes. A programme refers to one or more bitstreams that are used to represent the video and associated audio content. The use of digital techniques is frequently associated to the use of compression algorithms. High compression ratios are often achieved by using loss techniques while suffering minor decrease of picture quality. Thus, compression allows digitized video data to be represented in a much more efficient way and to broadcast programmes using only part of the bandwidth that would be necessary with raw video data. In broadcast applications, the simultaneous use of digital techniques and compression algorithms, such as the H.264/AVC standard, permits the reduction of transmission bandwidth while providing a service with a quality similar to or better than the previous analogue systems. Viewers receive video programmes from different video content providers via a transmission channel. Consider a basic broadcasting chain composed of a video coder connected to a video decoder via a multiplexer, a digital broadcast channel and a demultiplexer (Figure 1.1) ([29]). In a fixed multiplexing scheme, the channel bandwidth is divided into fixed parts between the different services so that the sum of the individual service's bandwidth is equal to or lower than the total available channel bandwidth.

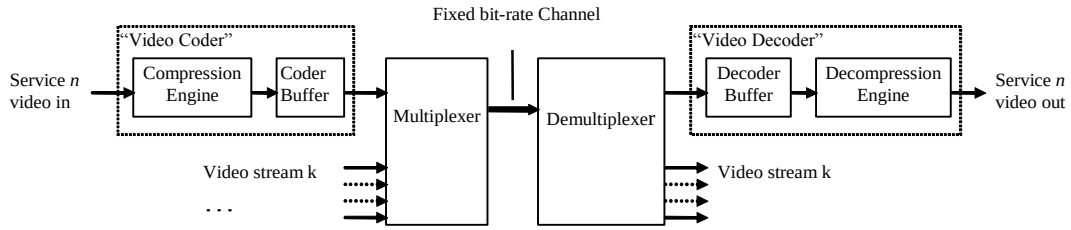


Figure 1.1 – A simplified digital broadcasting chain: encoder and decoder buffer diagram

The major issue for broadcasters is to select a criterion to divide bandwidth between the various services. To obtain constant video quality, encoders need to generate variable bit rate (VBR) data. Video signals representing varied programme types intrinsically have very diverse complexities. For example, a soccer game is much more complex to encode than an interview or a talk show. One way to solve this problem is to allocate different bit rates to each encoder based on the expected video complexity of the signal to be encoded. A larger fraction of the total bandwidth of the channel would be allocated to the soccer game programme compared to the talk show. This would result in a reduction of the picture quality variation between the different video programmes. One criterion for allocating bandwidth is based on the peak bit rate requirements. This method is inefficient, as total bandwidth would seldom be used. An

alternative is to allocate bandwidth among different programmes dynamically, adjusted over time, according to a criterion such as the relative complexity of each video programme. This process is named joint coding or statistical multiplexing [30] and can be described as follows [31]:

1. The bandwidth allocation process such that the aggregate instantaneous bit rate is not higher than the channel capacity, the minimum quality of service (QoS) requirements for all applications are met, and the quality is maximized for applications in the order of their importance. This allocation process depends on factors such as the complexity and relevance of each video programme.
2. The control required in cases where the aggregate instantaneous bit rate is greater than the channel capacity, so that it is possible to minimise the loss in QoS in many applications.

The way the bandwidth is allocated among programmes depends on the goal of the service. Several criteria have been used to allocate bandwidth within a joint coding process ([32],[33],[34],[35],[36],[37],[38],[39],[40],[41],[42],[43],[44],[45],[46],[47],[48],[49],[50],[51], [52],[53],[54]). Consider the case where bandwidth allocation is based on picture complexity. X_i denotes the picture complexity of programme i. A common allocation method is to allocate bandwidth, R_i for a video programme i, from the total bit capacity R, according to its coding complexity.

$$R_i = \frac{X_i}{\sum_{j=1}^{\text{Number\_Programmes}} X_j} R \tag{1.1}$$

References in literature mention a lower threshold for bit rate allocation below which the quality of compressed image drops abruptly ([55]). To solve this problem a minimum bit rate allocation $R_{\min}$ needs to be guaranteed for each programme i, and to allocate the remaining bits linearly, using a method similar to the Equation (1.2).

$$R_i = R_{\min_i} + \frac{X_i}{\sum_j X_j}\left[R - \sum_j R_{\min_j}\right] \tag{1.2}$$

Each video programme may have a different guaranteed minimum bit rate depending upon the anticipated overall complexity of the video transmitted through the channel and/or pricing of the channel to the providers of the video signals (Equation (1.2)).

A third method is to assign bandwidth proportionally to the value of a fee. The higher this value, the greater the fraction of the total bit rate of the transmission link allocated to that channel. Thus, better quality is obtained for the video programme with higher fee values. This approach would be represented by an equation similar to equation (1.1) where the complexity would be replaced by a weighting factor F_i associated with the pricing. A combined approach of pricing and complexity is also used and can be denoted as:

$$R_i = \frac{F_i X_i}{\sum_{j=1} F_j X_j} R \tag{1.3}$$

Another method is to guarantee a minimum allocation bandwidth per video programme and allocate the remaining bits using weighting factors that depend upon the anticipated overall complexity of the video signal transmitted over the channel and/or pricing of the channel to the provider of the video signals. This method can be formulated as follows.

$$R_i = R_{\min_i} + \frac{F_i X_i}{\sum_j F_j X_j} \left[R - \sum_j R_{\min_j} \right] \tag{1.4}$$

A common problem to all these methods is that usually the bit rate allocation for the next pictures is determined based on the complexity measures from preceding pictures. Thus, if there is a scene cut, the bits may not be enough to encode the new scene because the allocation was based on incorrect information. This problem can be avoided with a delay of one or two frames in the video encoding so that it is possible to detect a scene change. The ability to adapt the bit rate allocation of services sharing a multiplex is vital to a broadcaster for the following reasons:

1. Addition of new services by reducing the video bit rate of existing services within the multiplex and thus obtaining additional bandwidth capacity.
2. Removal of an inactive service, distributing the corresponding bandwidth and thus increasing the picture quality of the remaining video programmes.
3. Alteration of the video bit rates of services sharing a multiplex, maintaining the number of programmes within a multiplex, periodically (e.g. whenever each TV programme start) or continually (e.g. on a frame-by-frame basis) according to programme content. The aim is to maintain overall picture quality high by controlling the video bit rate allocation assigned to services according to the complexity of encoded material.

For digital television broadcasting applications, different environment scenarios can be foreseen for the joint video coding (Figure 1.2).

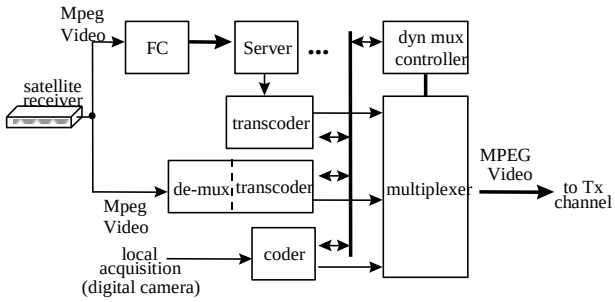


Figure 1.2 – Digital television environments

Figure 1.2 consists of a hybrid configuration, with local transcoders establishing the connection between remote encoders or previously stored video signals and the dynamic multiplexer, and with local video encoders for locally acquired programmes ([29],[50]).

Common to all though is the presence of stand-alone equipment implementing the joint video coding strategy. Scenarios will differ according to the location of video encoders relative to the multiplexer. For example, one interesting scenario for broadcasters should be to reassign bit rates while the multiplex is "on-air" and without causing the interruption of any service within the multiplex at a receiver.

Research in joint bandwidth allocation methods has been focusing on how to handle video quality among multiple video sources when the overall sum of bit rates on the channel is constant. Two approaches have been followed: how to obtain a uniform distribution of distortions among the different sources ([32],[33],[35],[36],[37],[38],[39],[40],[41],[42], [43], [44],[45],[46],[56]), and how to minimise the sum of distortions for each source ([34],[47],[57]).

In summary, the main contributions of this thesis are as follows: R-D Modelling of H.264/AVC using perceptual ways to assess the distortion; development of joint source coding algorithms for controlling a statistical multiplexing process of different streams into a fixed bandwidth channel that incorporates perceptual information; and study how joint coding results correlate with subjective quality assessment.

Additionally, the contributions of this Thesis include:

1. A survey of existing subjective and objective quality assessment metrics.
2. An overview of the main Digital Video Coding Standards.
3. A survey of the state-of-the-art HRD Models and Standards Rate Control Standard algorithms.
4. A bit rate variability study on H.264/AVC, coded using perceptual metrics.
5. Performance comparison between several Joint Video Coding Algorithms.
6. A two-pass video coding algorithm, incorporating perceptual metrics.

1.3 Outline of the Dissertation

The remainder of this dissertation is organized as follows:

Chapter 2 introduces the basic concepts of digital video quality, presents objective and subjective state-of-art methodologies and describes recent trends in video quality assessment. Besides providing the relevant background information in digital video quality assessment, the need for perceptual quality metrics is justified. It introduces two perceptual metrics: JND and SSIM.

Chapter 3 surveys the principal digital video coding standards. A detailed description of the new features of H.264/AVC standard is given.

Chapter 4 considers fundamentals of HRD models and standard rate control schemes, which is the basic core of this work. Although rate control is not part of the video coding standards, the algorithms developed during the standardisation process are an important reference in the video coding field.

Chapter 5 introduces rate control optimisation in H.264/AVC, and focuses on Rate-Distortion modelling for H.264/AVC. New Rate-Distortion models based on perceptual quality metrics are introduced and evaluated. In addition, the Bit Rate Variability-Distortion (VD) curve for H.264/AVC using perceptual quality metrics is proposed and analysed.

Chapter 6 presents the architecture and performance analysis of joint video coding using different complexity criteria.

Chapter 7 summarises the results and contributions of the Thesis.

Finally, the Annex section of this dissertation presents auxiliary information as well as additional experimental results for the various techniques developed in the thesis.

Chapter 2. Digital Video Quality and its Assessment

The assessment of video quality is a vital aspect for the effective designing, implementing, and monitoring of services by content providers. Digital video is being increasingly used in heterogeneous environments and applications, and thus subject to different types of distortions during acquisition, authoring, compression, transmission, and reproduction. One of the goals in the design of visual communication systems is to represent, broadcast and reproduce the information that the human eye can see and perceive ([58],[59],[60]). Failure to achieve this goal is a waste of resources (channel and terminal devices). Therefore, it is important to understand how pictures can be efficiently represented and assessed.

There are two emerging approaches for assessing subjective video quality ([61]). The first is Quality of Experience (QoE). The concept of QoE is recent and different definitions of QoE can be found in the literature ([62]). ITU-T Recommendation P.10/G.100 defines QoE as "the overall acceptability of an application or service, as perceived subjectively by the end-user" ([63]). The second approach is the concept of Quality of Perception (QoP), formally known as the user-perceived QoS (QoSE). ITU-T ([64],[65]) defines QoP (QoSE) as "a statement expressing the level of quality that customers/users believe they have experienced." Both QoE and QoP are user-centred approaches.

Quality assessment of digital video must be seen as an integral part of a complete compressed evaluation system. According to [66], a broad classification of the applications of quality assessment algorithms can be seen as applications that keep track of video quality for real-time applications, for assessing competing image and video processing algorithms, and to provide support to the optimisation of the design of image and video processing algorithms. The use of digital compression has expanded the types of distortions that can occur in a modern digital video system. Existing assessment methods for measuring video quality can be divided into two ([67],[68]): subjective testing (human observers provide their opinion on video quality) and the objective measurement's methods. This latter method is performed with the aid of instrumentation, either manually with humans reading a calibrated scale or automatically using a mathematical algorithm ([59]). A reliable way of assessing the quality of an image or video is the subjective evaluation, as human beings are the final receivers in most applications ([59]).

Nevertheless, it is a rather complex process (wide variety of possible methods and test elements) and generates too much variability in the results. This chapter gives a brief summary of the video quality assessment methodologies.

2.1 Subjective Assessment of Visual Quality

The production of audio-visual programmes' aims to satisfy human viewers so that their opinion of the video quality is relevant. Thus, regardless of the video quality metrics selected, it is rather important to ensure that it presents high correlation with subjective measurements. Indeed, controlled studies of how viewers assess quality must be regarded as the fundamental standard of performance for any video processing algorithm ([64]). Informal and formal subjective measurements have been used to evaluate system performance, from the design lab to the operational environment. This includes the psycho-visual assessment of video quality where video signals are viewed by human observers under restricted viewing conditions (fixed lighting, viewing distance, etc.) ([64]). In subjective tests, viewers (normally aged between 15 and 30 years) are asked to view a set of video sequences and give a quality score on a numerical or qualitative scale. The average value of the scores, for a given video sequence, is known as the Mean Opinion Score (MOS). Given that each human observer has distinctive interests and expectations when watching a video sequence, the subjectivity and variability of observer judgments cannot be totally removed ([69]). Subjective tests try to diminish these factors by unambiguous instructions, training and controlled environments. Nevertheless, a quality score is a noisy assessment that is described by a statistical distribution rather than a precise number. There is a broad diversity of subjective testing methods. The perceptual performance of subjects can be assessed by tools provided by psychophysics, beginning with visibility thresholds and just-noticeable differences (JND's), which are most appropriate for small impairments ([69]). Direct scaling methods have been standardised by the ITU through various recommendations ([70],[71],[72],[73]). The recommendations contain information regarding viewing conditions, criteria for the selection of observers and test material, assessment procedures, and data analysis methods. Selection of a specific subjective testing method depends on the application, quality variation, and the observer's duties ([69]).

Two broad measurement techniques exist. In *primary* or *explicit* measurements of picture quality, a group of viewers examines a set of pictures and makes subjective decisions on their quality. In *secondary* or *implicit* measurements, characteristics of standardised waveforms are objectively measured and the results are then converted to quality measures through previously established relations (e.g. waveform testing used in analogue transmission links). Primary methods are more useful if the distortions introduced by processing are complex in appearance,

as in some methods of digital coding. Both the primary and secondary methods assume that quality can be represented on a linear, one-dimensional scale. Multidimensional evaluations, which have been successfully applied in scaling speech quality, are also applied to television images ([74],[75]). Primary subjective evaluations are based on two broad methods, *Category-Judgments* methods (also referred to as *Rating-Scale* methods) and *Comparison* methods ([76]).

2.1.1 *Category-Judgments Methods*

In the case of category judgment's methods, viewers analyse a sequence of images processed according to a previously determined range of tests [58].

5 – Excellent	5 – Imperceptible	3 – Much Better
4 – Good	4 – Perceptible but not annoying	2 – Better
3 – Fair	3 – Slightly annoying	1 – Slightly better
2 – Poor	2 – Annoying	0 – Same
1 – Bad	1 – Very annoying	-1 – Slightly worse
		-2 – Worse
		-3 – Much worse
(a)	(b)	(c)

Table 2.1 – Quality and Impairment Ratings Commonly Used

After viewing the video sequences, viewers classify each video sequence according to a previous selected category such as the overall quality or the visibility of impairments (Table 2.1 a) and Table 2.1 b)). The subject's response usually depends on many factors such as the highlight luminance, contrast ratio, ambient room light, picture size, viewing distance, experience and motivation of the subjects, and the range of the picture material. Proper experimental design is necessary to avoid biases in the results due to factors such as the order of presentation of the processed pictures. In addition, variability among subjects may be minimised by using “expert” subjects who are familiar with principles of television, particularly the visual appearance of impairments. However, experts are also more sensitive to imperfections in a picture and may not be representative of general viewers.

2.1.2 *Comparison Methods*

In the comparison method, the subject compares an impaired or distorted test picture with a reference picture to which a standard-type impairment (e.g. white noise) has been added ([58],[60]). The comparison may be on two monitors arranged side by side or on one monitor where both pictures are displayed sequentially in time. Impairment is added to the reference picture until both pictures appear to the subject to be of equal quality. The amount of this added impairment can be under the subject’s control or, alternatively, pictures with variable amounts of impairments can be computed in advance, stored and displayed in a given sequence. The

comparison between the test picture and the reference picture can usually be done reliably when the two types of distortions are visually similar, e.g. additive noise of different spectral characteristics. The distortion of the test picture can be assigned a quality (or category) using the previous category-judgment tests on the impaired reference picture ([58],[60]). In a variation of this method, the subject uses a comparison rating scale, for example Table 2.1 c), to compare test pictures with different levels of distortion with a reference picture. The subject replies to the question "how much better or worse is the test picture compared to the reference picture?". The resulting data is subsequently processed to determine the "point of subjective equality" between the distorted test and impaired reference picture.

2.1.3 *Subjective Standardisation Efforts*

Subjective testing is used for quality assessments, system performance under optimum conditions, and impairment assessment under non-optimum performance due to transmission limitations ([67],[68]). In a digital television system, picture quality is not a constant over time. Picture quality is a function of the complexity of the programme material. Considering this time-varying property and the number of new impairments, the number of assessment methods has grown in recent years. Other factors have been the focus of study such as viewing conditions, choice of observers, scaling method to score the opinions, reference conditions, signal sources for test scenes, timing of the presentation of the various test scenes, selection of a range of test scenes, and analysis of the resulting scores ([64],[69]). Selection of the parameters for each of these elements depends on the requirements of the television system. Typically, an assessment process starts with the selection of non-expert observers, by means of an examination of their visual capabilities. They then view a series of test scenes for about 10 to 30 minutes in a controlled environment and are asked to score the quality of the scenes in one of a variety of manners.

ITU has standardised methodologies for the subjective assessment of the visual quality of television pictures, in ITU-R Rec. BT.500 ([70],[71]), and multimedia systems in ITU-T Rec. P.910 ([72]). Experimental setup and viewing conditions diverge in the two recommendations ([77]). Recommendation ITU-R BT.500 has been used in studies and field trials to assess the subjective quality of television pictures in large formats. It defines several subjective testing methods for television system that in some cases include two or more scoring procedures ([70],[71]). A vast range of basic test methods has been used in television assessments. However, specialised methods should be used to deal with specific assessment problems. Table 2.2 presents an overview of representative assessment problems and of the methods typically used to deal with these problems. A description of the methods will be provided in this section.

Assessment problem	Method used	Description
Measure the robustness of systems (i.e. failure characteristics)	DSIS	Rec. ITU-R BT.500, § 4
Measure the quality of systems relative to a reference	DSCQS	Rec. ITU-R BT.500, § 5
Measure the quality of stereoscopic image coding	DSCQS	Rec. ITU-R BT.500, § 5
Measure the fidelity between two impaired video sequences	SDSCE	Rec. ITU-R BT.500, § 6.4
Compare different error resilience tools	SDSCE	Rec. ITU-R BT.500, § 6.4

Table 2.2 – Selection of test methods - Rec. ITU-R BT.500 ([71])

Recommendation P.910 has been targeted at multimedia applications such as videoconferencing, usually reduced picture's formats such as QCIF and CIF, and the new varieties of display screens such as LCD.

Depending on whether a reference video is used in the subjective tests, the suggested test methods can be regarded as a double-stimulus or single stimulus scheme ([71]). In the case of double-stimulus methods, the observer can compare the reference and processed videos so that they can be more precise in their judgments. The disadvantage is that it requires more time to perform the assessments compared with the single stimulus methods. In this latter case, viewers only watch the video and then they have to assess its quality. This process allows more opinions to be gathered in a limited time, thus increasing the accuracy of the trial. These two assessments methods perform rather similarly in terms of precision ([71]). As mentioned previously, a short description of different subjective assessment schemes will be provided in order to give some insight. More information is available in [70],[71],[72].

Double Stimulus Impairment Scale (DSIS) – also referred to as Degradation Category Rating (DCR), where multiple reference-scene, degraded-scene pairs are shown to observers. First, an unimpaired reference is displayed, followed by the impaired version of same picture or sequence. Observers assess the amount of impairment in the test video compared with the reference video, on an overall impression scale of impairment, with the following qualifiers: imperceptible, perceptible but not annoying, slightly annoying, annoying, and very annoying ([71],[77]). While DSIS has been used with limited ranges of impairments, it is more appropriately used with a full range of impairments [71].

Double Stimulus Continuous Quality Scale (DSCQS) - Multiple scene pairs with the reference and degraded scenes are randomly shown to observers Figure 2.1 in a pseudo-random order.

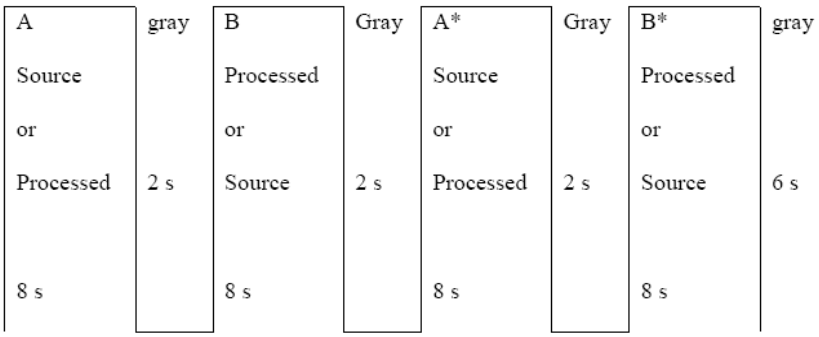


Figure 2.1 – Presentation structure of test material for DSCQS ([71])

Each scene pair is individually rated. Scoring is performed on a continuous five-grade quality scale, ranging from bad to excellent, as Figure 2.2 shows ([71],[77]). A grading scale is generated, in pairs, containing the double assessment of each test pair.

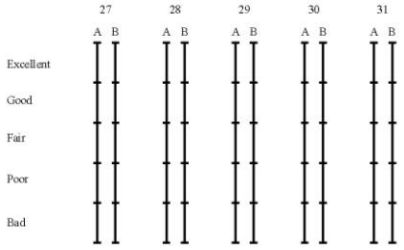


Figure 2.2 – DSCQS grading scale ([71])

The duration of the trials should not exceed half an hour with explanations, and preliminaries included. The drawback of DSCQS is its redundancy that limits the extension of assessment sequences ([71]).

Single Stimulus Method (SS) – this method is also named Absolute Category Rating (ACR). It is a category judgement where the video sequences are display one at a time without explicit references ([71]). After each presentation, observers evaluate the quality of the sequence shown on a category scale.

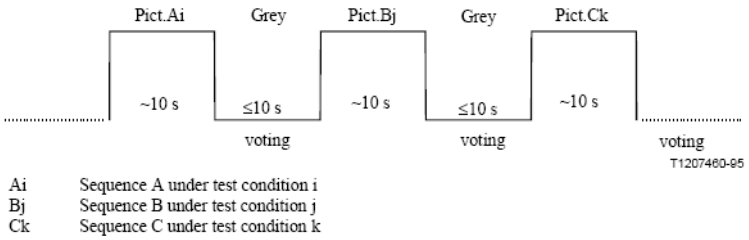


Figure 2.3 – Stimulus presentation in the ACR method ([71])

Figure 2.3 indicates the time pattern for the sequence's presentation. Each test sequence should last about 10 seconds, and the viewer should rate it in a time less than or equal to 10 seconds.

In general, three types of SS methods have been used in television assessments: adjectival categorical judgement methods, numerical categorical judgement methods, and non-categorical judgement methods ([71]). In the first method, viewers evaluate test sequences according to one of a set of categories. These categories are defined typically in semantic terms such as Quality or Impairment. The other two methods consist of an 11-grade numerical categorical scale (useful if a reference is not available) and a continuous scale with no numbers or a large range, e.g. 0 – 100 ([71]).

ACR is an effective method that allows a large number of test sequences to be assessed in a short period of time ([72],[78],[79],[80]). To prevent the assessment values from being affected by differences in the video content used in test sessions, a new method was defined that includes a reference version of each test sequence. The Absolute Category Rating with Hidden Reference (ACR-HR). It was formalised in ITU-T Recommendation P.910 ([72]). During the data analysis, assessment results are determined by computing the difference in scores between the test sequence and its corresponding (hidden) reference. This procedure is known as "hidden reference." Results are expressed as DMOS (differential quality scores).

Simultaneous Double Stimulus for Continuous Evaluation (SDSCE) - in general, the time duration of video sequences under assessment is limited to 10 seconds ([71],[77]). However, this is not representative of the duration of sequences in a real service. The SDSCE approach was intended to assess longer sequences. Thus, viewers observe two sequences, reference and test version, side-by-side at the same time. Assessment is then performed by monitoring the differences between the two sequences and scoring the fidelity of the test video using a slider. In order to produce relevant statistical analysis, the time duration of a test sequence should be at least two minutes. The drawback of this approach is that viewers have to switch their attention between two images from time to time.

Single Stimulus Continuous Quality Evaluation (SSCQE) - a programme is continuously evaluated over a long time period, 10 to 20 minutes, without a source reference. Data is sampled from a continuous scale every few seconds. Scoring is a distribution of the amount of time a particular score is given.

Subjective Assessment Methodology for Video Quality (SAMVIQ) – the methods that have been presented so far have been developed by ITU. SAMVIQ was specified by the European Broadcasting Union (EBU) ([81]). Quality assessment is carried out scene after scene, including an explicit reference, a hidden reference and various algorithms (codec's). SAMVIQ is based on random presentation of the video sequences. Video sequences are shown in multi-stimulus form, so that the observers can choose the order of tests and alter their scores. Due to the presence of

an explicit reference, observers can directly compare the impaired sequences among themselves and against the reference, thus giving the score accordingly. It is to be noted that SAMVIQ was the first method that has been developed mainly for multimedia, incorporating the differences on how multimedia content and television are watched.

All these methods are extensively used and are generally considered trustworthy for video quality assessment. Nevertheless, the methods differ in several aspects such as the use of explicit or hidden reference sequences, the use of high and low anchors, the structure of the test organisation, the display of one or two videos simultaneously and the way viewers score test files such as continuous scoring compared to a single evaluation.

Parameter	DSIS	DSCQS	SSCQE	SDSCE	SAMVIQ
Explicit reference	yes	no	no	yes	yes
Hidden reference	no	yes	no	no	yes
High anchor	no	yes	no	no	hidden reference
Low anchor	no	yes	no	no	yes
Sequence length	10 s	10 s	5 min	10 s	10 s
Two simultaneous stimuli	no	no	no	yes	no
Presentation of test material	I: once II: twice in succession	II: twice in succession	once	once	several concurrent (multi-stimuli)
Quality evaluation	once	once	continuous	continuous	once
Possibility of changing vote before proceeding	no	no	no	no	yes

Table 2.3 – A summary of the subjective quality methods ([82])

Kozamernik in [82] has summarised the different characteristics of the subjective methods (Table 2.3). This type of assessment presents several advantages. Results are expressed as a scalar mean opinion score (MOS), produced for both conventional and compressed television systems, and it works well over a broad range of still and motion picture applications. This type of test presents various weakness: the large variety of existing methods, the meticulous setup and control required, the high number of selected observers and screenings that can result in increased complexity and a very time-consuming process. A final note refers to the conceptual dissimilarities between the significance of the quality scale descriptors. It is known to vary between linguistic groups, cultural groups, and between individuals, to a non-negligible extent ([83]). Results are analyzed on the assumption that conceptual discrepancy is uniform.

2.2 Objective Assessment of Visual Quality

Subjective quality assessment is the most important method of estimating the perceived quality. Nevertheless, it is time-consuming, quite complex, costly, and requires singular assessment

facilities to generate reliable and reproducible test results. Thus, much effort has been spent on designing algorithms that are able to characterise subjective quality and predict viewers' opinion. In this section, classifications and characteristics of existing objective video quality metrics and assessment models will be addressed, and state-of-the-art metrics will be discussed.

2.2.1 *Objective Quality Assessment Models*

To deal with different requirements, distinctive types of quality estimation and prediction models have been developed for various domains of application and range of system or service conditions ([84],[85]). A universal quality model that can be applied in all conditions does not exist. It is possible to categorise models according to different criteria such as the application goals, input parameters, network components and configuration under evaluation, predicted quality features, etc ([84],[85],[86]).

Objective quality assessment models are used in different contexts such as planning, lab-testing, and monitoring ([84]). In the first case, the goal is to predict the perceived quality of networks/systems' services prior to implementation. In the second case, the aim is to estimate in the laboratory the quality of networks/systems' services as the equipment is being developed. Finally, in the last case, the focus is on the quality prediction of networks/systems' services that are already operational. In general, objective quality assessment methodologies can be categorised into five types. These are media-layer models, parametric packet-layer models, parametric planning models, bitstream layer models, and hybrid models (Table 2.4) ([85],[87]).

<table>
<tr><th></th><th>Media-layer model</th><th>Parametric packet-layer model</th><th>Parametric planning model</th><th>Bitstream layer model</th><th>Hybrid model</th></tr>
<tr><td>Input information</td><td>Media signal</td><td>Packet header information</td><td>Quality design parameters</td><td>Packet header and payload information</td><td>Combination of any</td></tr>
<tr><td>Primary application</td><td>Quality benchmarking</td><td>In-service non-intrusive monitoring (e.g. network probe)</td><td>Network planning, terminal/ application designing</td><td>In-service nonintrusive monitoring (e.g.,terminal-embedded operation)</td><td>In-service nonintrusive monitoring</td></tr>
<tr><td>Audio</td><td>ITU-R BS1387</td><td rowspan="3">ITU-T P.NAMS [IPTV]</td><td rowspan="3">ITU-T G.1070 [videophone]
ITU-T G.OMVS [IPTV]</td><td rowspan="3">ITU-T P.NBAMS [IPTV]</td><td>—</td></tr>
<tr><td>Video</td><td>ITU-T J.144 [SD]
ITU-T J.vqhdtv [HD]
ITU-T J.mm**[PC]</td><td>ITU-T J.bitvqm [IPTV]</td></tr>
<tr><td>Multimedia</td><td>(ITU-T J.148)</td><td>—</td></tr>
<tr><td>Speech</td><td>ITU-T P.862</td><td>ITU-T P.564</td><td>ITU-T G.107</td><td>—</td><td>ITU-T P.CQO</td></tr>
</table>

Table 2.4 – Objective quality assessment models and corresponding standards and ongoing projects in ITU ([87])

Media-layer quality models use the knowledge of HVS to determine the quality of video. There are two steps to represent human visual processing: psychophysical models (low-level visual

information processing) and cognitive models (high-level functions) ([87]). Media-layer quality models do not require a-priori knowledge of the system under testing (e.g. codec type). Thus, they can be use in the assessment of unknown systems such as codec comparison/optimisation).

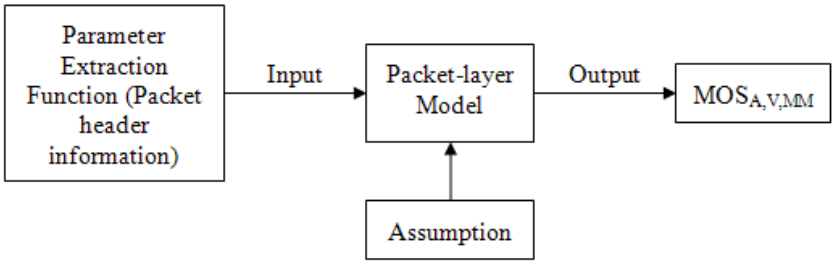


Figure 2.4 – Packet-based Model ([84])

In a parametric packet-layer model, quality is predicted based only on packet-header information and performance criteria are provided in terms of quality estimation (Figure 2.4) This type of model allows real-time and detailed quality monitoring. In a parametric planning model, quality planning parameters for networks and terminals are used as input. Thus, prior information regarding the system under evaluation is needed. ITU-T Recommendation G.107, frequently named the E-model, is an example of a parametric planning model. It has been extensively used as a network planning tool for the PSTN and VoIP services) ([87]).

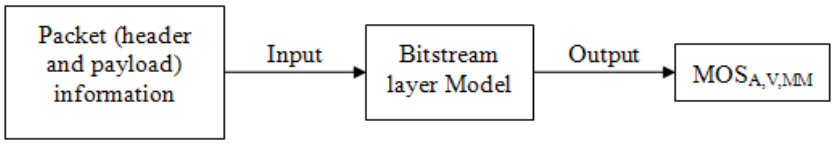


Figure 2.5 – Bitstream-layer Model ([84])

Parametric packet-layer models are very efficient in terms of processing requirements as they do not analyse the payload information (Figure 2.5). On the other hand, it is more difficult for them to predict quality as no information regarding video content is available. Media-layer models have access to the video content information but processing video content requires additional computational power.

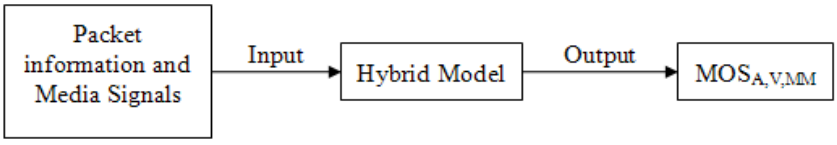


Figure 2.6 – Example of Hybrid Model ([84])

The bitstream-layer model occupies an intermediate position between these two types of models. They use encoded bitstream information and packet-layer information that is used in parametric packet-layer models. The hybrid model is a combination of the technologies discussed so far, such as packet information, bitstream or even decoded video as input (Figure

2.6) ([77]). It is effective in terms of exploiting as much information as possible to estimate video quality ([88]).

2.2.2 *Visual Perception*

There are many theories of visual perception. They frequently differ in their style, content, and the level at which perception is described and explained. Some of these differences arise because they adopt starting points. Some theories have favoured possible physiological mechanisms, which could mediate certain aspects of seeing. Others, by contrast, have insisted that human perception is described by psychology. These differences also determine the relative importance attached to bottom-up versus top-down factors in perceiving terms.

2.2.2.1 Sensory Thresholds

Making sense of (or perceiving) the environment is a major achievement. Roth describes the process of visual perceptions as being "the means by which information acquired from the environment via the sense organs is transformed into experiences of objects, events, sounds, tastes, etc." ([89]). Thus, perception stands for processing specifically sensory information instead of including the thoughts and behaviours resulting from detected stimulus ([90]).

A distinction is sometimes drawn between perception and sensation, with sensation referring to the experience of the basic not interpreted information presented to the sense organs ([91]). It has been argued that sensation occurs before perception, but it is more realistic to assume that they generally overlap in time. Traditionally, some believe that the processes involved are so complex that there is a little value in trying to divide them up neatly into sensation and perception ([91]).

To understand how the senses operate in protecting the human person and enriching their life requires an appreciation of *sensory processes*. Psychologists have been working towards distinguishing between *sensation* and *perception* ([92]). The starting point for both processes is a *stimulus*, a form of energy (for example, light waves) that affects sensory organs. *Sensation* is the process that detects stimuli from one's body or environment. *Perception* is the process that organises sensations into meaningful patterns. Visual sensation lets you detect the black marks on this page; visual perception lets you organise the black marks into letters and words. Sensation depends on specialised cells called *sensory receptors*, which detect stimuli and convert their energy into neural impulses. This process is called *sensory transduction* ([91]). If a stimulus remains constant in intensity, one will gradually stop noticing it. This tendency of sensory receptors to respond less and less to an unvarying stimulus is called *sensory adaptation*.

Sensory adaptation let humans detect potentially important changes in their environment while ignoring unchanging aspects of it.

2.2.2.2 Psychophysics

One of the key questions in this area is how much change in light intensity must occur for us to notice it? *Psychophysics* has been developed as the study of the relationship between the physical characteristics of stimuli and the corresponding psychological response to them, addressing a question like this ([60],[91]).

The minimum amount of stimulation that a person can detect is called the *absolute threshold*, or *limen*. Because the absolute threshold for a particular sensory experience varies, psychologists operationally define the absolute threshold as the minimum level of stimulation that can be detected 50 percent of the time when a stimulus is presented repeatedly. Besides detecting the presence of a stimulus, one must be able to detect changes in its intensity. The minimum amount of a change in stimulation that can be detected is called the *difference* (or relative) *threshold*. Formally, the definition of difference threshold is the minimum change in stimulation that can be detected 50 percent of the time by a given person. Weber referred to the difference threshold as the *just noticeable difference* (JND) ([60]). He found that the amount of change in intensity of stimulation needed to produce a JND is a constant fraction of the original stimulus.

$$\frac{\Delta I / I}{I} = k \tag{2.1}$$

This became known as *Weber's law*. In Equation (2.1) ΔI is the JND in the intensity of the stimulus, I is the original intensity (the standard stimulus) and k is a constant for a particular sensory modality ([60]). Other Weber fractions have been established in every sensory modality. Fechner's work based on Weber's work, led him to conclude that to produce incremental, *arithmetic* steps in sensation, the physical stimulus must grow *geometrically*. Thus, he reformulated Weber's law as

$$\text{Sensation} = K \log (\text{Stimulus intensity}) \tag{2.2}$$

where K represents a weighting constant for each modality (and includes the relevant Weber fraction), and the logarithmic multiplier represents the exponential or geometric growth in stimulus intensity required to yield successive JND ([93]).

Fechner's contribution in this area was two-fold. First, he reinforced the idea that scales relating sensory sensitivity and stimulus intensity sensibility are compressed; magnitude of sensation does not grow in a linear form with stimulation. Secondly, he initiated the development of psychophysical methods for the exploration of sensation and perception. The essence of these

methods is that stimuli are presented in a systematic manner and that the observer's task is simplified by requiring him or her to make one of a restricted set of responses.

Researchers, inspired by Fechner's work, have devised *signal-detection theory* (SDT) ([94]), which assumes that the detection of a stimulus depends on both its intensity and the physical and psychological state of the individual. One of the most important psychological factors is *response bias* – how ready the person is to report the presence of a particular stimulus. Signal-detection researchers study four kinds of reports that a subject might make in response to a stimulus. A *hit* is a correct report of the presence of a target stimulus. A *miss* is a failure to report a target stimulus that is present. A *false alarm* is a report of the presence of a target stimulus that does not exist. A *correct rejection* is a correct report of the absence of a target stimulus.

	Response 'Yes'	Response 'No'
Signal present	Hit	Miss
Signal absent	False alarm	Correct rejection

Table 2.5 – Stimulus/response matrix

With the application of SDT to the psychological phenomena, a new model of the human perceiver was proposed. When applied to vision, the new model assumes the presence of a physiological process in the visual system, the value of which varies randomly according to a Gaussian distribution over time. However, in the approaches described in this section, the observer is viewed as an essentially passive receiver of stimulation. A paradigm shift occurred with the emergence of a different kind of model: an intuitive statistician following decision rules in a manner analogous to those used in testing scientific hypotheses.

2.2.2.3 Vision Models

Different assessment metrics for image and video quality are based on HVS models, in particular, in psychophysics. In psychophysics, it is possible to approach viewing by describing the visual system as a "black box" (an input-output system where visual stimuli are the input, and sensations the output) ([91],[93]). Thus, the transfer function defines the visual system. This approach could be used in the video coding field, primarily with the *visibility* of coding impairments. This section shall briefly describe experiments in threshold vision that are used to evaluate the visibility thresholds of some stimuli.

Contrast Sensitivity Functions

The response of HVS depends on the relation of its local variations to the surrounding luminance rather than on the absolute luminance ([59]). Contrast is a measure of the relative

variation of luminance. This property is known as the Weber-Fechner law and may be denoted as:

$$C^W = \frac{\Delta L}{L} \tag{2.3}$$

The threshold contrast is also defined as the minimum contrast necessary for an observer to detect a change in intensity. This value depends on stimulus characteristics, especially colour as well as its spatial and temporal frequency ([95]). Contrast sensitivity is defined as the inverse of the contrast threshold. The contrast sensitivity function (CSF) describes the sensitivity of the HVS to different spatial and temporal frequencies that are present in the visual stimulus. The CSF chart shows the contrast sensitivity for all spatial frequencies ([95]). Human observers are most sensitive to intermediate frequencies (~4 - 8 cpd) and less sensitive to lower and higher frequencies. In CSF measurements, Michelson contrast is defined as the contrast of periodic stimuli with varying frequencies and can be expressed by the following equation ([59]):

$$C_M = \frac{L_{\max} - L_{\min}}{L_{\max} + L_{\min}} \tag{2.4}$$

where $L_{\max}, L_{\min}$ are the luminance extreme values of the pattern. The CSF shows the observer's window of visibility. Points below the CSF are visible to the observer as it displays even higher contrasts than the threshold level.

Masking

Masking is an important aspect of the HVS in modelling the interactions between different images components present at the same spatial location ([59],[60],[91]). Masking refers to the fact that the presence of one image component (called the mask) will decrease/increase the visibility of another image component (called the test signal). The mask generally reduces the visibility of the test signal in comparison with the case where the mask is absent. However, sometimes the opposite effect, facilitation, occurs when a stimulus that is not visible by itself can be detected due to the presence of another. Spatial masking effects are usually quantified by measuring the detection threshold for a target stimulus when it is superimposed on a masker with varying contrast ([96]). The visibility threshold of a particular stimulus depends on many factors. When restricting to achromatic stimuli, the principal factors are the average background (spatially and temporal constant) luminance level against which the stimulus is presented, the suprathreshold luminance changes (in space and time) near the test stimulus in space and time, and finally, the spatial shape and temporal variation of the stimulus.

It is useful to think, in a quality assessment context, how distortion or coding noise may be masked by the original image or sequence acting as background masking. This might explain why similar coding artefacts are disturbing in certain regions of an image while they are hardly noticeable elsewhere. Typically, the masking effect is strongest when the mask and the test signal have comparable frequency content and orientations. Most quality assessment methods incorporate one model of masking or the other, while some incorporate facilitation as well ([97],[98]). Several vision models propose that masking occurs only between stimuli located in the same channel, between channels with distinct orientation ([99]), between channels of different spatial frequency, and between chrominance and luminance channels ([100],[101]).

In the literature, two main forms of temporal masking have been identified: shot scene and the temporal contrast sensitivity function ([60],[91]). A shot scene occurs when there is a major change in the video programme content. This change induces a remarkable increase in the masking levels for a period of up to 100ms after a shot scene ([102]). Temporal masking is thus very important for interframe coding. However, temporal masking is difficult for at least two reasons: television cameras integrate the image of any object on the target (motion-related blurring and resolution loss), and the perception of a moving object heavily depends on whether or not the object is tracked by the eye. Temporal masking can occur before and after a discontinuity, respectively “backward masking” and “forward masking” ([103]). In the first case, it may be explained as the result of the variation of the latency of the neural signals in the visual system as a function of their intensity ([103]). Temporal facilitation, which is the opposite of temporal masking, can occur at low-contrast discontinuities ([104]).

Pooling

The pooling operation concerns the task of determining a single measurement of quality, or a decision regarding the visibility of the artefacts, from the outputs of the visual streams ([59]). It is not totally understood how the HVS performs pooling. In fact, there is no firm experimental evidence that the mathematical assumption general employed in the literature is a good description of a pooling mechanism in the HVS ([105],[106]). Clearly pooling involves cognition, where a perceptible distortion may be more annoying in some areas of the scene (such as human faces) than at others. Pooling can be computed as follows:

$$E = \left\{ \sum_{l} \sum_{k} \left| e_{l,k} \right|^{\beta} \right\}^{1/\beta} \tag{2.5}$$

where $e_{l,k}$ is the normalised and masked error of the kth coefficient in the lth channel, and β is a constant typically with a value between 1 and 4 (values approximate to the number 2 produce

goods results ([107]), β=2.4 Lubin ([108]), β=2 Teo and Heeger ([98])). This form of error pooling is commonly called Minkowski error pooling. Minkowski pooling may be performed over space (index k) and then over frequency (index l), or vice-versa, with some non-linearity between them, or possibly with various β exponents. A spatial map showing the relative importance of each image region may also be used to provide spatially variant weighting to the different $e_{l,k}$ ([109],[110],[111],[112]). Minkowski pooling [Equation (2.5)] is used in several quality assessment metrics to pool the error signal from the different frequency and orientation selective streams, as well as across spatial coordinates.

General Framework of Perceptual Image Quality Metrics

Most of the proposed quality metrics that model the HVS share a similar structure as they are based on models of low-level HVS processing (optics, the retina, the lateral geniculate nucleus, and the striate cortex) [113]. They share a common paradigm: to determine the strength of the errors between the reference and the distorted signals in a perceptually meaningful way.

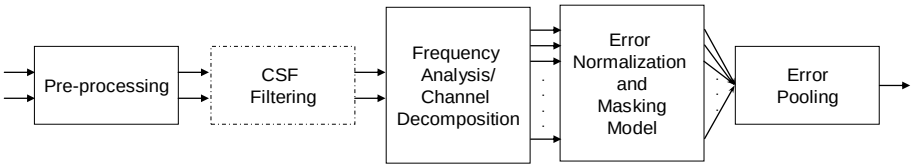


Figure 2.7 – Perceptual framework

Figure 2.7 shows a general perceptual based quality assessment framework based on HVS modelling ([59],[60]). Most of the quality assessment algorithms that model the HVS can be explained with this framework, although they may differ in the specifics.

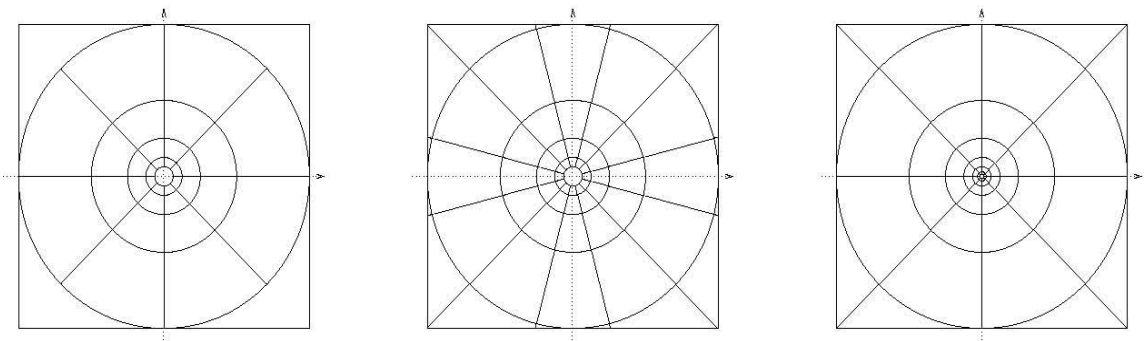
Pre-processing usually involves operations such as alignment, colour space conversion; calibration and display model adaptation. The alignment guarantees a point-to-point correspondence between distorted and reference signals. Colour space conversion provides a representation of the signals in a linear perceptual space colour (example: CIE Luv and CIE Lab [114]). [115] is recommended for a deeper analysis of standardised colour spaces. In a digital system, it is necessary to convert digital images into physical luminance through non-linear transformations (images from different devices that may have gone through different transformations such as gamma correction) to be able to display images. Finally, an exact model of the display device may be employed as viewers can only see what the display can reproduce ([116]).

CSF may be implemented using linear filters that approximate the frequency responses of the CSF or are implemented as weighting factors for channels following frequency analysis. A model of the CSF for luminance, proposed by Mannos and Sakrison ([117]) is given by:

$$CSF(f) = K_0(1 + K_1 f)e^{-(a.f)^{\alpha}} \tag{2.6}$$

where K_0, K_1, a and α depend on parameters such as mean luminance, temporal frequency and orientation ([118],[119]). The CSF or other similar curves have been extensively used to design quantisation matrices for Hadamard, DCT ([120],[121],[122],[123]), or lapped orthogonal ([124]) transforms. These approaches have been extended to incorporate image dependent components into the JND thresholds ([109],[125]). The CSF has been modelled for video quality assessment as simple temporal filters ([110],[126],[127]). An overview of the usage of visual models in compression is presented in [128].

Frequency analyses consist of a hierarchy of filters that decompose the image into several components or channels (usually called sub-bands) with distinct spatial frequencies and orientations. The channels allow the division of the visual stimulus into different spatial and temporal sub-bands. Several models have been proposed in the literature varying from sophisticated channel decompositions, such as the 2D Gabor or Cortex transform, to simpler transforms such as the wavelet transform or even the DCT. One reason for selecting a simpler transformation is their suitability for certain types of applications rather than their accuracy in representing the cortical neurons. A sophisticated approach, such as 2D Gabor functions, well represents the cortical receptive fields. Nevertheless, the Gabor decomposition is difficult to compute and lacks some of the mathematical characteristics that are considered necessary for good implementation, such as invertibility, reconstruction by addition, etc. Watson has proposed the cortex transform ([129]), which have equivalent profiles as 2D Gabor functions, but it is easier to implement. This balance between implementation requirements and conformity regarding HVS has led to the proposal of different models in the literature: Watson ([129]), Daly ([130]), Lubin ([97],[108]), and Teo and Heeger ([98],[131],[132]). Some examples of such decompositions are shown in Figure 2.8. Each axis varies from $-u_s/2$ to $u_s/2$ cycles per degree, where u_s is the sampling frequency.



a) b) c)

Figure 2.8 – Frequency decomposition for Watson (a), Daly (b) and Lubin (c) models

Figure 2.8 a) shows the Cortex transform proposed by Watson. The Cortex transform consists of two classes of filters applied sequentially that decompose the image, first into separate radial frequency bands, and then into different orientation bands. It attempts to approximate the radial and orientation selectivity of HVS. Several variations of the Cortex transform have been proposed: Figure 2.8 b) and Figure 2.8 c) shows, respectively, Daly ([130]) and Lubin ([108]) proposals. One of the differences in Daly's proposal is the six orientation bands. Lubin ([108]) proposed the Laplacian pyramid ([133]) to decompose the image into seven radial frequency bands and the steerable filters ([134]) to decompose each pyramid level into four different orientations. All the radial filters of these decompositions have octave bandwidths. Teo and Heeger adopted the steerable pyramid transform by Simoncelli et al. ([135]), which also has octave radial bandwidths and six orientation bands.

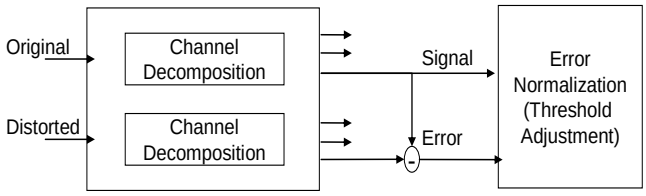


Figure 2.9 – Implementation of masking effect for a channel

Error normalisation and masking are usually implemented within each channel, normally through a gain-control mechanism that weights the error signal in a channel by a space-varying visibility threshold designed for that channel (Figure 2.9) ([136]).

The energy of the original signal (or both the original and the distorted signals) in the surrounding area of each point determines the visibility threshold adjustment. Another important factor is the HVS sensitivity on that channel in the absence of masking effects (also called base-sensitivity). As mention previously, masking is generally defined as any interference between two or more visual signals or stimuli that results in a modification of their visibility ([59],[60],[91]). Several independent masking phenomena exist.

Luminance masking corresponds to the process when the luminance of the original image signal masks the variations in the distortion signal ([109]). One method to account for luminance masking, also called light adaptation, is by including a modification of the base sensitivities. The luminance masking adjustment is a function of the local luminance. A possible simplification is to consider that it is independent of the k sub-band index.

Spatial masking effects are usually quantified by measuring the detection threshold for a target stimulus when it is superimposed on a masker with varying contrast ([96]). For example, the edges in images can mask signals of much greater amplitude than a region of near-constant intensity ([58]). This phenomenon has a physiological explanation: lateral inhibition

([137],[138],[139]). The effect of spatial masking is rather limited and is concentrated in a location very close to the edge (only a few pixels in natural images). Furthermore, the masking effect of edges is only important when the masked signal has the same orientation as the edge.

The term contrast masking is used to denote the case where both the target and the masker have the same frequency and orientation ([131]). The term texture masking is used to refer to the general case. Most of the quality metrics only proposes to model the Contrast Masking ([140]). Initial psycho visual tests have been performed on sinusoidal patterns of a single frequency and a single orientation. Extension of these results to introduce frequency and orientation dependencies has been successful ([99],[131],[136]). These models correctly predict the contrast detection threshold for a target signal, usually a sinusoid or Gabor patch, in the presence of a masking signal (a sinusoid or Gabor patch of different frequency, phase, and contrast).

Existing perceptive models of human vision based on Physiology corroborate results from physiology studies ([119]). The retina divides the visual stimulus existing in an image into components with the following characteristics: the position in the visual field (in the image), the spatial frequency (in the Fourier domain: the amplitude in polar coordinates), and orientation (in the Fourier domain: the phase in polar coordinates).

One perceptual channel can only be excited by the component of a signal that has similar characteristics. Furthermore, it appears that signals with the same components take the same visual pathways from the eye to the cortex. Masking models that only take account of the interactions inside one channel of the visual pathways are called intra-channel models. The models that take into account the interaction between different channels are called inter-channel models ([136]).

Two different explanations are used for pattern masking ([136]). In the first explanation, the mechanism detecting the target has a nonlinear, compressive response. The mask activates this mechanism, and pushes its response into the compressive range. The differential between responses to mask alone and to target plus mask is thereby reduced, and threshold elevated ([96],[141]). In the second explanation, the mask inhibits the target detection mechanism, either directly or through other mechanisms. Several psychophysical models, inspired by research of the response properties of single visual neurons in a primary visual cortex ([142],[143],[144]), have been proposed. These models incorporate both mechanisms within a process of contrast gain control ([98],[99],[145]). In this case, contrast gain control is a mechanism that serves to keep neural responses within their permissible dynamic range while retaining the information conveyed by the pattern of activity over the neural ensemble. In the normalisation model of Heeger ([144]), each neuron has an accelerating non-linearity. However, it is also inhibited

divisively by a pool of responses of other neurons. In the psychophysical model of Teo and Heeger that is closely based on this cortical normalisation theory, masking occurs through the inhibitory effect of this normalizing pool. Foley's model of masking also incorporates a divisive inhibitory pool. The final step, error pooling, is the process of combining the error signals (normalised by the sensitivity thresholds) in different channels (computed for each spatial frequency, orientation band and each spatial location) into a single value or a distortion matrix.

2.2.3 *Classifications of Objective Quality Metrics*

According to Jacobson, image quality can be defined as 'the subjective impression formed in the mind of the observer relating to the degree of excellence exhibited by an image' ([146]). This subjective impression is then the result of a variety of contribution factors such as the physical properties of the observed image, the observer's experience and future expectations, environmental conditions, context and pictorial content ([72],[147],[148]). The question that should be formulated is "How do we measure this subjective impression?" The existence of a functional relationship between the subjective impression of image quality and some aspect of the physical image that is observed can be assumed ([149]). Thus, it is possible to construct suitable psychophysical tests that measure subjective quality using a panel of observers. Nevertheless, this is a complex and time-consuming task that is hard to perform. So, the coding community typically favours objective methods that are usually more consistent and less complex to implement. A large number of measures for objective image quality can be found in the literature. Usually, based on the image properties, three categories of image quality measurements can be identified: image distortion measures (evaluation of physical properties in terms of differences in pixel values or differences in spectral power distribution), image fidelity measures (relate physical measures to the visual properties of the image), and image distortion measures (relate physical measures of an image (or a set of images) to the perceived subjective quality of that image).

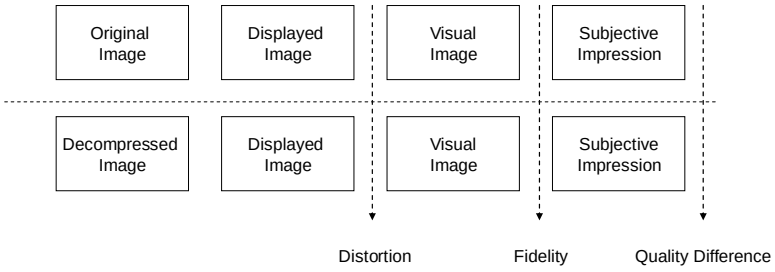


Figure 2.10 – Image quality measurements and their location in a digital imaging system

Figure 2.10 illustrates the three categories and their position in a digital imaging system. Distortion measurements assess the physical differences between images requiring a test or

control. Although useful, it cannot tell whether a particular physical difference is visible to an observer. Two commonly used metrics are the Mean Square Error (MSE) and Peak Signal-to-Noise Ratio (PSNR). Formally, the MSE is defined as follows ([150],[151]):

$$MSE = \frac{\sum_{i=1}^{M}\sum_{j=1}^{N}\left| f(i,j) - F(i,j)\right|^2}{M \times N} \tag{2.7}$$

where f(i,j) is the original video component at pixel (i,j), F(i,j) is the impaired video component at pixel (i,j), M is the picture width, and N is the picture height. PSNR is obtained 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 as follows

$$PSNR = 10\log_{10}\left(\frac{255^2}{MSE}\right) \tag{2.8}$$

This type of metrics fails to predict the HVS perception because they take no account of where errors occur in the image, not every change in an image is noticeable or leads to distortion, no error is visually important. Other distortion measures have been published, some attempting to include the effects of the display and viewing systems. Examples include the mean square error after non-linearity (MSENL) and distortion contrast measures ([152]), and a measure based on differential Sobel filtering ([153]). However, various studies ([152],[153],[154]) have shown that these distortion measures correlate poorly with perceived image quality. In particular, the addition of a constant value to every pixel produces a large response from a distortion measure but has a very small visual effect, a slight increase in image brightness.

Fidelity measures focus more in the visual threshold, the point at which the visual system detects a difference between the stimuli. Since it is impractical to measure the 'visual image', fidelity measurements operate by modelling the visual system to compute visual images based on the physical properties of the original scene and the associated viewing conditions. This enables the visibility of artefacts to be predicted under different display and viewing conditions.

Fidelity metrics only predict visibility. The typical output is in the form of a just noticeable difference (JND) scaled visibility map ([130]). This means that, in general, the metrics are measuring the threshold of visual detectability, and using a multiplier of that threshold to indicate the magnitude of the fidelity loss. These types of metrics are less useful for predicting image quality since the proper scaling of supra-threshold distortions requires additional inputs from the cognitive system ([147]). Increased distortion, or decreased fidelity, does not necessarily involve a decrease in quality. Fidelity metrics are also particularly complex to

implement; the many observer-specific parameters result in conflicting implementations, making it harder for them to gain widespread acceptance and making it difficult to transfer results between tests.

The third category is image quality (or true image quality). This describes the subjective impression of the excellence of the image. It is concerned with both the threshold points (whether a process causes a perceptible change in quality) and the supra-threshold magnitudes (whether a perceptible change is an increase or a decrease and its magnitude). It has expressed the difference between quality and fidelity more simply as the difference between the visibility of a factor and the degree to which that factor is annoying ([155]). Since quality is a subjective perception based on a variety of attributes, it can be assessed either as a comparison between two (or more) images, or independently with only one image. Using a single image, memory, experience and expectations become more dominant, and the relationship between physical properties and perceptions of quality becomes harder to measure or predict ([147],[149]).

According to [69],[77],[156], the measurement of video distortions in a video communication system can be distinguished in two ways: data metrics and picture metrics. Data metrics evaluate the fidelity between original and processed videos, without taking into account the content of the video under analysis. Typical metrics in this category are the MSE and the PSNR. The main benefit of this type of metrics is that they are easy to compute, while the disadvantage is that the visual importance of the pixels is disregarded.

In the case of Picture metrics, the quality assessment is oriented towards the content of the video under analysis. The impact of distortions and content on perceived quality is taken into consideration. Comparing with the Data metrics method, these metrics are closer to the human perceived quality. The Picture metrics can be classified in two groups, namely a *vision modelling approach,* also referred to in the literature as the psychophysical approach, and an *engineering approach* ([60]). The vision modelling approach is based principally on HVS. This type of approach seeks to include human vision characteristics that seem to be important to picture quality, like contrast sensitivity and pattern masking, colour perception, applying models and data from psychophysical experiments. According to Winkler ([69]), metrics based on HVS date back to the 1970s and 1980s. More recent metrics in this category are the Visual Differences Predictor (VDP) by Daly ([130]), the Sarnoff JND (just noticeable differences) metric by Lubin ([97],[108],[157]), or van den Branden Lambrecht's Moving Picture Quality Metric (MPQM) ([110]).

In the engineering approach, overall quality is predicted based on the extraction and analysis of special features, such as contours, or artefacts in the video, like block artefacts, introduced by a

specific video processing step, compression technology, or transmission link. This type of metrics does not automatically ignore the attributes of the HVS, as they frequently take into account psychophysical effects too. However, the conceptual basis for their design is the image content and distortion analysis, instead of fundamental vision modelling. This approach has become more popular in recent years ([66]). One example is Wang *et al.*'s Structural Similarity (SSIM) index ([69]). It determines the mean, variance and covariance of small areas within an image and combines the measurements into a distortion map. More information about SSIM will be provided in subsequent sections.

A different way to classify metrics is based on the amount of information about the reference video required and available to predict the video quality. These approaches are classified in full-reference (FR), no-reference (NR) and reduced-reference (RR) categories ([60],[77],[158],[159]). They will be briefly reviewed.

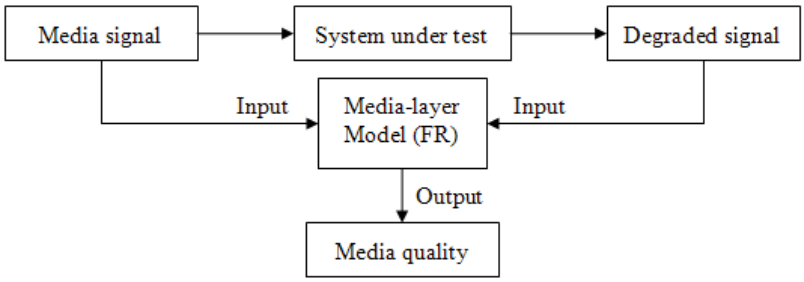


Figure 2.11 – FR Diagram Block ([84])

In the FR methods (Figure 2.11), an image-by-image comparison is performed between the reference video and the test video. Most of the objective metrics are categorised in this group ([160]). MSE, PSNR and HVS-based metrics belongs to this class ([77]). It is assumed that the entire reference video is accessible, and that exact spatial and temporal alignment between the two videos sequences exists. This is a major constraint for applications since the reference video is not always available or transmission sometimes suffers frame drops.

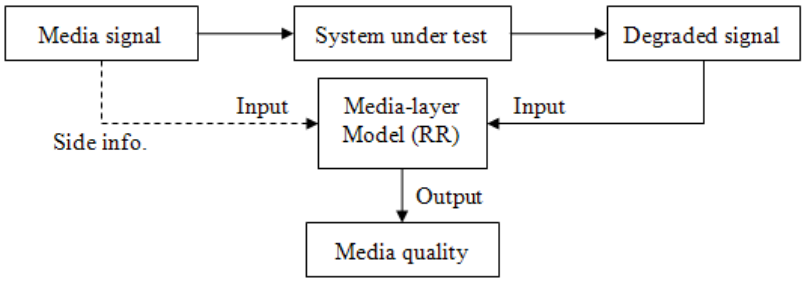


Figure 2.12 – RR Diagram Block ([84])

In NR methods (Figure 2.12), video quality is assessed without access to any information other than the impaired signal. Thus, no prior knowledge about the reference video is needed. In

addition, the mentioned alignment issues are avoided. This type of metrics has been concentrating the attention of the industry ([67]). For example, in a wireless cellular network, video signals suffer a wide variety of compression, channel, and processing distortions. The possibility of using FR methods in real-time in this type of scenarios is extremely remote as access to the reference sequence is very difficult. This would be an ideal scenario for NR methods. The principal problem of NR metrics is how to separate distortion from content. Human observers are able to make this distinction in a simple way. They are capable of assessing the quality of a video just by seeing the test video and without viewing its reference ([161]). It is one of the "Holy Grails" of the image-processing field ([67]). As HVS knowledge is restricted, work in this area is limited ([162]). Usually, NR metrics has to make presumptions about the video content and the distortions. In general, NR metrics based their assumptions through the extraction of features such as blockiness, blurring, quantisation, or ring artefacts ([77],[161],[162]).

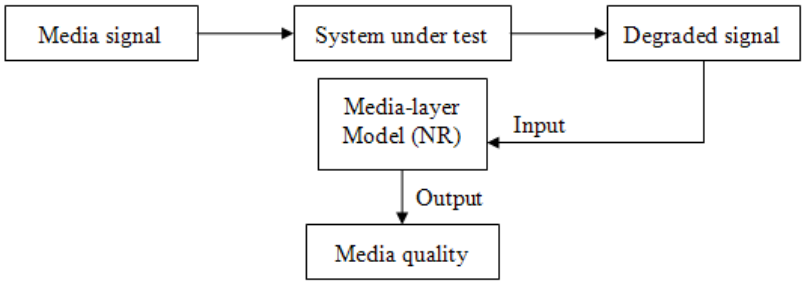


Figure 2.13 – NR Diagram Block ([84])

In RR metrics (Figure 2.13), algorithms work without the reference signal. Instead, they have access to the test video signal along with additional (side) information and/or extracted features, such as the localised spatio-temporal activity information, the amount of motion or spatial detail, detected edge locations, or embedded marker bits to estimate the distortion of the channel ([77],[163],[164]). Assessment is made based on the comparison of those features that need to be aligned. This is a more flexible requirement compared with the reference alignment. The RR metrics are a compromise between FR and NR metrics.

These three distinct approaches have different operational uses. In general, FR metrics are used for offline video quality measurement. One typical example is the adjusting of a video codec, where it is possible to control environment, and where a good analysis of the video is needed. NR and RR metrics are better tailored for monitoring in-service video systems. In these cases, it is essential to obtain real-time assessment ([77]). Comparing the distinctive metrics, it has been claimed that the best results are reached when the FR metrics are used ([162]).

2.2.4 *Objective Standardisation Efforts*

In the beginning of the nineties, the fast progress of digital video technologies created a major challenge to the performance measurement field: the need for a new measurement methodology to assess the performance of digital video systems. With this objective, the ANSI Technical Subcommittee T1A1 approved a series of three standards: ANSI T1.801.01 ([165]), ANSI T1.801.02 ([166]), and ANSI T1.801.03 ([158]). These standards aimed to categorise key impairments in video signals and recommend ways for measuring those impairments ([158],[165],[166],[167]).

In the case of ANSI T1.801.3, the goal is to present methods for measuring principal video impairments in a practical system (mostly teleconferencing applications) ([158]). Different types of parameters known as scalar, vector and matrix parameters are described. Parameters typically measure a spatial (SI) or temporal (TI) perceptual property of the sequence ([168]).

The Spatial perceptual Information feature, SI, describes the activity of image edges or spatial gradients. A digital video system can add edges (e.g. edge noise) or reduce edges (e.g. blurring). SI is based on the Sobel filter. First, the luminance component of each frame, at a time n, $\mathrm{Y_n}$,is filtered with a Sobel filter. Then the standard deviation over the pixels of the Sobel-filtered frame is computed. This procedure is repeated for each frame in the video sequence, and SI corresponds to the higher standard deviation value. The Sobel filter is implemented by convoluting two 3 × 3 kernels over a frame of a video sequence, and applying the square root of the sum of the squares to the results of these convolutions. SI_h and SI_v are the outputs from filtering a frame by a 3x3 horizontal and vertical edge-detecting Sobel filter, respectively. Let $Y(i, j, n)$ denote the pixel of the input image n at the ith row and jth column. $\mathrm{SI}_v(i, j, n)$ will be the result of the first convolution and is given by:

$$\mathrm{H}_v = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix}, \mathrm{H}_h = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} \tag{2.9}$$

$$\mathrm{SI}_v(i, j, n) = Y(i, j, n) \otimes \mathrm{H}_v \tag{2.10}$$

$$\begin{aligned} \mathrm{SI}_v(i, j, n) = & -1 \times Y(i-1, j-1, n) - 2 \times Y(i-1, j, n) - 1 \times Y(i-1, j+1, n) \\ & + 0 \times Y(i, j-1, n) + 0 \times Y(i, j, n) + 0 \times Y(i, j+1, n) \\ & + 1 \times Y(i+1, j-1, n) + 2 \times Y(i+1, j, n) + 1 \times Y(i+1, j+1, n) \end{aligned} \tag{2.11}$$

Similarly, $\mathrm{SI}_h(i, j, n)$ will be the result of the second convolution and is given by:

$$\mathrm{SI}_h(i,j,n) = Y(i,j,n) \otimes \mathrm{H}_h \tag{2.12}$$

$$\begin{aligned}\mathrm{SI}_v(i,j,n) = & -1\times Y(i-1,j-1,n) - 2\times Y(i-1,j,n) - 1\times Y(i-1,j+1,n) \\ & + 0\times Y(i,j-1,n) + 0\times Y(i,j,n) + 0\times Y(i,j+1,n) \\ & + 1\times Y(i+1,j-1,n) + 2\times Y(i+1,j,n) + 1\times Y(i+1,j+1,n)\end{aligned} \tag{2.13}$$

The computations are performed for all $2 \le i \le N-1$ and $2 \le j \le M-1$, where N is the number of rows, and M is the number of columns. After computing SI_h and SI_v the next step is to compute SI_r and SI_θ using the following expressions:

$$\mathrm{SI}_r(i,j,n) = \sqrt{SI_h^2(i,j,n) + SI_v^2(i,j,n)} \tag{2.14}$$

$$\mathrm{SI}_\theta(i,j,n) = \operatorname{atan}\left(\frac{SI_h(i,j,n)}{SI_v(i,j,n)}\right) \tag{2.15}$$

For each frame, SI_{stdev} is computed as the standard deviation of SI_r

$$\mathrm{SI}_{mean}(n) = \frac{1}{total\ number\ of\ pixels}\sum_i\sum_j \mathrm{SI}_r(i,j,n) \tag{2.16}$$

$$\mathrm{SI}_{\mathrm{var}}(n) = \frac{1}{total\ number\ of\ pixels}\sum_i\sum_j \left(\mathrm{SI}_r(i,j,n) - \mathrm{SI}_{mean}(n)\right)^2 \tag{2.17}$$

$$\mathrm{SI}_{stdev}(n) = \sqrt{\mathrm{SI}_{\mathrm{var}}(n)} \tag{2.18}$$

SI_{stdev} is the principal statistical metric for spatial information that is used in determining ANSI T1.801.3 spatial parameters ([158],[167]). Parameters based on spatial information can be understood as clues regarding added or lost edges in the destination video sequence compared to the original video sequence. Added edges are caused from impairments such as tiling, error blocks, and noise. Lost edges can result from impairments like blurring. SI is the maximum observed value of SI_{stdev} in all the video frames within the video sequence.

The Temporal perceptual Information, TI, characterises the activity of temporal differences or gradients between consecutive frames. Increased levels of motion in adjacent frames will result in higher TI values. An encoding system can add motion (e.g. error blocks, jerkiness, or noise) or reduce motion (e.g. frame repeats). TI is based on the difference (motion) feature that is the difference between the luminance pixel values of two adjacent frames, $\mathrm{TI}(i,j,n)$. $\mathrm{TI}(i,j,n)$ as a function of time (n) can be defined as:

$$\mathrm{TI}(i,j,n) = \mathrm{Y}(i,j,n) - \mathrm{Y}(i,j,n-1) \tag{2.19}$$

Similar to Spatial Information, SI, Temporal Information, TI, is determined as the higher value over time of the standard deviation of $\mathrm{TI}(i,j,n)$.

$$\mathrm{TI}_{mean}(n) = \frac{1}{total\ number\ of\ pixels}\sum_i\sum_j \mathrm{TI}(i,j,n) \tag{2.20}$$

$$\mathrm{TI}_{\mathrm{var}}(n) = \frac{1}{total\ number\ of\ pixels}\sum_i\sum_j \left(\mathrm{TI}(i,j,n) - \mathrm{TI}_{mean}(n)\right)^2 \tag{2.21}$$

$$\mathrm{TI}_{stdev}(n) = \sqrt{\mathrm{TI}_{\mathrm{var}}(n)} \tag{2.22}$$

Figure 2.14 displays the spatial activity of the fourth frame of a video sequence of football, before and after being encoded by an H.264/AVC codec (CBR 256kbps, IPPP GOP1, JM rate control ([169],[170]), more information in the next Chapters. It also displays the image difference between the SI of the original and degraded version. To guarantee a clear visualisation of the effect of lost edges, the display range was adjusted by a factor of two. One of the steps in an H.264/AVC encoding algorithm is quantisation. Usually, high-frequency components of the image may be reduced or suppressed. As high-frequency components are associated with edges in the frame, the impact of quantisation can be observed. It is possible to verify that some edges are missing. The intensity of this effect depends on the nature of the video and the level of quantisation.

Figure 2.14 – Football SI (from left to right: original image filtered with Sobel edge filter, encode image filtered with Sobel edge filter, image difference)

Figure 2.15 illustrates the TI feature. To generate an impaired image an H.264/AVC codec with the previous settings of SI example was used. A football sequence is characterised by fast motion with a high level of variation between neighbouring frames. In the sequence, American football players try to advance the ball down the field by throwing the ball, attempting to catch the ball in order to advance down the football field with the ball as far as they can go. The video camera is focused on the ball, follows its movements through a pan (horizontal movement in

which the camera moves left to right). Figure 2.15 shows a greater pixel magnitude, centring in on the players that are going to start running.

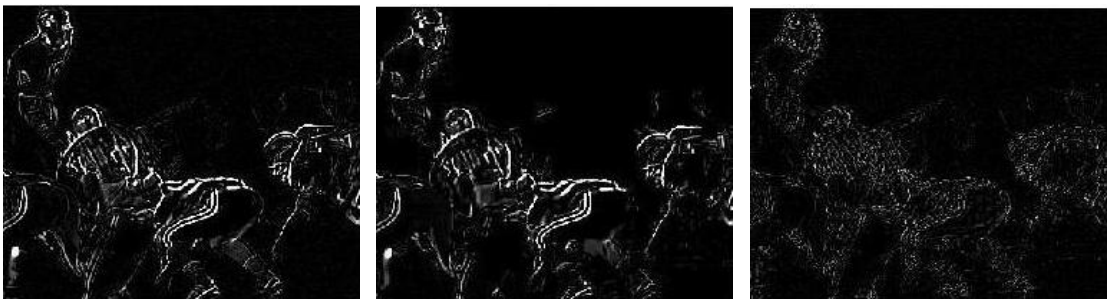


Figure 2.15 – Football TI (from left to right: original image, encoded image, image difference)

IEEE Broadcast Technology Society Subcommittee on Video Compression Measurements started an approach to the problem of video quality assessment. The goal was to develop a scale of video impairment and unit of measurement to characterise video distortion from both a perceptual and engineering perspective ([171]). It was suggested that this study attempt to define a scale of video impairment in terms of several measurements of the just-noticeable difference (JND). The outcome of this recommended study was never reported. However, in reaction to the perceived necessity in the industry for standards, T1A1.1 decided in February 2001 to develop a series of four Technical Reports (TRs). These reports, approved in October 2001, provide full disclosure of VQMs currently used by industry, and an extensible framework into which properly documented VQM can be incorporated and quantitatively related to already disclosed VQMs. The quantitative relationship is established using a subjective dataset and set of videos (distorted and undistorted) that were viewed in the subjective-rating experiments. The dataset that was used is the VQEG set ([78],[79]). The first TR in the framework (TR A1 [172]) covers methods for specifying the accuracy and cross calibration of the video quality metrics. The second TR in the framework (TR A2 [173]) covers normalisation methods (e.g. spatial registration, temporal registration, and gain/level offset calibration). The third (TR A3 [174]) covers specification of one video quality metric that is commonly used by industry, namely peak-signal-to-noise-ratio (PSNR). The fourth (TR A4 [175]) specifies a JND-based VQM utilizing the full reference technique.

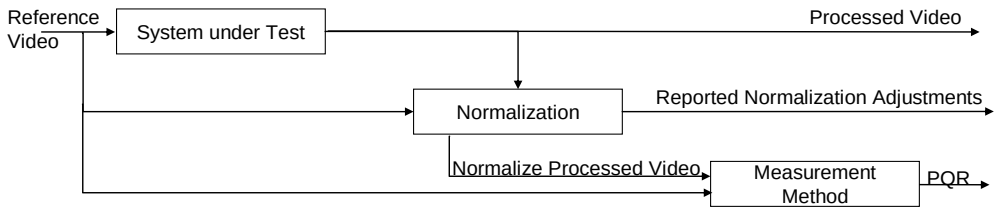


Figure 2.16 – System block diagram

A last method is a double-ended measurement identified as the PQR (Picture Quality Rating) (Figure 2.16). The PQR method specified in TR A4 is based on processing an 8-bit digital component video as defined by ITU-R Recommendation BT.601. Due to the perceptual nature of the measurement, various compression methods can be accommodated (MPEG, NTSC, PAL, etc.). In addition, the transmission system may include a concatenation of compression methods or be a simple pass-through for the evaluation of a codec.

Video Quality Experts Group - VQEG

To create a framework for the evaluation of new objective methods for video quality evaluation the Video Quality Experts Group (VQEG) was formed in 1997 as an informal subgroup of experts from three groups, ITU-R SG11, ITU-T SG9 and ITU-T SG12.

The main goal of the VQEG is to provide input to the relevant standardisation bodies responsible for producing international recommendations regarding the definition of an objective Video Quality Metric (VQM) in the digital domain ([78],[79]). Under the VQEG umbrella, four groups were formed: Independent Labs and Selection Committee (ILSC), Classes and Definitions, Objective Test Plan, and Subjective Test Plan. The test plan defines the procedure for evaluating the performance of objective video quality models as submitted by the ITU. It allows for assessing and comparing correlations between objective and subjective methods ([78],[79]).

During a first phase, FR-TV Phase I, from 1997 to 2000, VQEG designed and implemented extensive subjective and objective test plans. The goal was to evaluate a number of perceptually based proponent algorithms for the FR method, including the commonly used PSNR ([158]), focused on out-ofservice quality testing ([78],[79]). It opened a call for the submission of proposals for objective assessment models. It required that all models should accept as input an impaired sequence and the original sequence. With this input, the model was assumed to produce one distinctive value that correlated with the value achieved in the subjective tests. A set of test sequence was selected. Evaluation sessions were performed on a total of 287 viewers to gather the subjective data, while nine objective models were assessed using statistical analysis on three aspects of their capacity to predict subjective assessment of video quality, namely prediction accuracy, prediction monotonicity and prediction consistency. Prediction accuracy is the ability to predict the subjective quality ratings with low error (two metrics were employed, specifically the variance-weighted regression correlation and the non-linear regression correlation [78]). Prediction monotonicity corresponds to the degree to which the model's predictions agree with the relative magnitudes of subjective quality ratings (Spearman's rank order correlation [78]). And finally, prediction consistency is the degree to which the

model maintains prediction accuracy over the range of video test sequences (outlier ratio [78]). This results of these tests, named as full reference television (FR-TV) Phase I, yielded inconclusive results (none of the proposed models statistically out-performed any of the others nor were they statistically better than PSNR) ([78]). This gave VQEG increased motivation to look for more results that are consistent. One of the main accomplishments of this process was the particular data set gathered to support future development of objective models.

In the period 2001-2003, VQEG performed FR-TV Phase II ([176],[177]). Many FR bidders made improvements to their original algorithms. The VQEG FR-Phase II tests focus on secondary distribution of digitally encoded television quality video (525-line video and 625-line video). Each experiment covered a vast range of quality, so the assessment criteria were better able to determine statistical differences in model performance. Furthermore, it contained a wide set of typical content (motion complexity, spatial detail, colour, etc.) and normal video processing conditions to evaluate the ability of models to perform consistently over a very broad set of video content. Several models of the VQEG FR-TV Phase II performed substantially better than the conventional PSNR. Therefore, four models (BTexact, CPqD, NTIA, and Yonsei University/Radio Research Laboratory) were proposed for inclusion in the normative section of the Recommendation ITU-R BT.1683 ([178]). However, it also recognised that objective video quality is still a developing technology and ITU-R WP 6Q encouraged the proposers of the recommended models to collaborate towards this goal. Substantial progress in the development of FR metrics was achieved from the test results of both phases, especially for quality metrics that targeted MPEG-2 coding distortions.

On March 2004, the first official meeting of the Joint Rapporteurs Group (JRG) on Multimedia Quality Assessment (MMQA) took place during the ITU SG12 meeting. The JRG on MMQA brings together experts in audio quality assessment from SG12 and experts in video quality assessment from SG9. The JRG MMQA is responsible for co-coordinating activity associated with the development and testing of objective perceptual quality models for multimedia services. In September 2009, VQEG finished an assessment of metrics for multimedia applications that were focused on broadband Internet and mobile video streaming, at bit rates below four Mbps, with smaller frame sizes (QCIF, CIF, VGA). This project, VQEG Multimedia Phase I (MM-I), was used to validate full-reference, reduced-reference, and no-reference objective models ([79],[179]). Based on this report, two new standards for multimedia quality assessment were published, namely ITU-T Rec. J.247 ([180]), which defines four FR models (OPTICOM, Psytechnics, Yonsei University, and NTT), and ITU-T Rec. J.246 ([23]), which defines one new RR model for multimedia, that of Yonsei University ([79],[181]). These standards have been intended for telecommunications services broadcast at four Mb/s or less.

In August 2009, VQEG completed the RR and NR tests for SDTV (RRNR-TV) ([182]). These tests were an extension to the tests on FRTV, Phase I and II ([78],[176]). H.264/AVC and MPEG-2 codecs were used in the tests. All NR models were withdrawn. The final report describes the performance of seven RR models. The ITU decided that the accuracy of some RR models was sufficient to justify standardisation.

Project	FRTV_I	FRTV_II	MM_I	NRRRTV	HDTV
Timeline	1997-2000	2000-2003	2004 -2008	2000-2009	2004 -2010
Proponents	9 + PSNR	6 + PSNR,	25 + PSNR	7 + PSNR	8 + PSNR
Focus	FR TV videos	Secondary distribution of digital encoded TV	Mobile and broadband internet communication	SDTV	HDTV application
Model Types	FR	FR	FR, RR, and NR	NR (withdrawn) and RR	FR, RR, and NR (withdrawn)
Subjective Test	DSCQS, 5 scale DMOS	DSCQS, 5 scale DMOS	5-scale ACR-HR DMOS	5-scale ACR-HR DMOS	5-scale ACR-HR DMOS
Evaluation Metrics	4 metrics (after polynomial or logistic mapping)	7 metrics (after logistic mapping)	Pearson Correlation, RMSE, Outlier Ratio (after polynomial mapping)	Pearson Correlation, RMSE, Outlier Ratio (after polynomial mapping)	Pearson Correlation, RMSE (after polynomial mapping)
HRCs Considered	16 HRCs, MPEG2 H.263 bit rate 768kbps/ 50 mbps analogue videos, 625/50 and 525/60, with transmission errors.	10 HCR for 625/50 14 HCR for 525/60, MPEG2 H.263 bit rate 768 kbps/ 5 mbps	VGA/CIF/QCIF H.264/H.263/ MPEG2/MPEG4. compression artefacts, transmission error, prepost- processing effects, live network conditions, interlacing problems, bit rates 16kpbs- 4 mbps, variable frame rates	MPEG2 H.264, transmission error, bit rate 1 - 5.5 mbps	1080i/p, MPEG2 H.264, compression artefacts, transmission error, pre- and postprocessing, frame rate 25/30fps bit rate 1-30Mbps,

Table 2.6 – Summary of VQEG projects

In 2004, VQEG started a project to evaluate models for HDTV ([183]). The H.264/AVC and MPEG-2 video codec were used. Furthermore, distortion types like transmission error, pre- and post- video processing were included, and the bit rate was ranged from 1 Mbps up to 30 Mbps. The test plan comprised FR, RR and NR objective video quality models. Models were submitted in 2009 and VQEG's Final Report was approved in 2010 ([183]). All NR models were withdrawn. The ITU determined that the accuracy of one FR model and one RR model

were satisfactory to support standardisation. Table 2.6 presents an overview of completed VQEG projects.

Currently, VQEG have several active projects: 3DTV, HDTV Phase II, Hybrid Perceptual/Bitstream, JEG-Hybrid, Multimedia Phase II, and Quality Recognition Tasks (QART). The 3DTV Project aims to evaluate 3DTV subjective video quality. The HDTV Phase II project proposes to complete a second round of validation of HDTV objective video quality models using the available datasets from HDTV Phase I. The aim is to foster the development of more accurate models. The Joint Effort Group (JEG) Hybrid project aims to produce a robust Hybrid Perceptual/Bit-Stream model. The JEF is a recent idea of VQEG. It offers an alternative collaborative action instead of the competitive traditional process. Results will increase knowledge regarding objective quality assessment for video using bitstream and decoded video information. The Multimedia Phase II project is researching topics regarding audio and visual subjective quality. Finally, the QART project's goal is to study the effects of resolution, compression and network effects on the quality of video used for recognition tasks in order to foster the standardisation of the quality assessment model for image recognition.

Observing VQEG outcomes, from the different closed and active projects, it is possible to extract some conclusions regarding the tendencies in the development of objective video quality assessment. Although there has been substantial progress, it is still impossible to replace, in all the cases, subjective quality metrics for objective quality metrics. Relevant progress has been accomplished in the development of FR metrics. More research is needed in NR and RR algorithms. Soon research will focus on quality assessment of audio-visual content.

2.2.5 *Just Noticeable Distortion (JND)*

With the fast development of visual applications there is increasingly significant demand to incorporate perceptual characteristics into applications to improve performance. In [128],[184] Jayant described a major concept of perceptual coding, namely, just noticeable distortion (JND). The perfect JND should provide, for each video signal that is going to be encoded, a threshold level of the error visibility, below which reconstruction errors are imperceptible. According to Jayant ([128],[184]), the JND profile of an image is a function of local signal properties. Thus, it derives from various masking effects in the HVS. The mapping of these characteristics to a unique value requires an efficient perceptual model derived from wide subjective testing.

A good JND model can considerably improve the performance of video encoding applications. Some techniques for computing JND have been proposed, from the discrete cosine transform (DCT) and wavelet domain ([109],[185],[186],[187]), to the image-domain ([187],[188]).

Approaches in literature mainly focus on computing the quantisation step sizes in sub-band JND ([109],[185],[186],[187],[188]). A small number of experiences have been made in non-standard video encoding systems. Chou and Li ([187]) have presented a model for luminance components only, and the model has been used in rate control techniques (e.g. H.263 [189] and MPEG-4 [190]. The following expressions describe the perceptual model for estimating the JND profile, according to Chou and Li ([187]):

$$\mathrm{JND} = \max\left\{ f_1(\mathrm{bg}, \mathrm{mg}),\, f_2(\mathrm{bg}) \right\}, \tag{2.23}$$

$$f_1(\mathrm{bg}, \mathrm{mg}) = \mathrm{mg} \cdot \alpha(\mathrm{bg}) + \beta(\mathrm{bg}), \tag{2.24}$$

$$f_2(\mathrm{bg}) = \begin{cases} T_0 \times \left(1 - \left(\frac{bg}{127}\right)^{0.5}\right) + 3, if \le 127 \\ \gamma \times (bg - 127) + 3, \text{otherwise} \end{cases} \tag{2.25}$$

$$\alpha(\mathrm{bg}) = \mathrm{bg} \times 0.0001 + 0.115, \tag{2.26}$$

$$\beta(\mathrm{bg}) = \lambda - \mathrm{bg} \times 0.01 \tag{2.27}$$

where f_1 and f_2 model the spatial masking effect and luminance adaptation respectively, bg and mg correspond to the average background luminance and the maximum weighted average of luminance differences around the pixel at (x,y) coordinates, respectively, and T_0, γ and λ are 17, 3/128 and 1/2, respectively. All the parameters were empirically determined by fitting the model with subject test results ([187]). The average background luminance, bg, is calculated by the following equation

$$bg(x, y) = \frac{1}{32} \sum_{i=1}^{5} \sum_{j=1}^{5} p(x - 3 + i, y - 3 + j) \cdot B(i, j) \quad \text{for } 0 \le x \le H \text{ and } 0 \le y \le W, \tag{2.28}$$

where $B(i, j),\ \text{for } i,\ j = 1, 2, \ldots, 5$ is a weighted low-pass operator, shown in Figure 2.17, and $p(x, y)$ denotes the pixel at (x,y).

1	1	1	1	1
1	2	2	2	1
1	2	0	2	1
1	2	2	2	1
1	1	1	1	1
B				

Figure 2.17 – Matrix B for determining average background luminance

To obtain $mg(x,y)$ across a pixel $p(x,y)$ it is necessary to compute the weighted average of luminance changes around the pixel $p(x,y)$ in four directions ([187]):

$$mg(x,y) = \max_{k=1,2,3,4}\left\{\left|grad_k(x,y)\right|\right\} \tag{2.29}$$

with

$$grad_k(x,y) = \frac{1}{16}\sum_{i=1}^{5}\sum_{j=1}^{5} p(x-3+i, y-3+j)\times G_k(i,j) \tag{2.30}$$

Figure 2.18 presents the four operators, $G_k(x,y)$, that are used to perform the computation, where the weighting coefficient decreases as the distance away from the central pixel increases.

0	0	0	0	0		0	0	1	0	0
1	3	8	3	1		0	8	3	0	0
0	0	0	0	0		1	3	0	-3	-1
-1	-3	-8	-3	-1		0	0	-3	-8	0
0	0	0	0	0		0	0	-1	0	0
G1						G2				
0	0	1	0	0		0	1	0	-1	0
0	0	3	8	0		0	3	0	-3	0
-1	-3	0	3	-1		0	8	0	-8	0
0	-8	-3	0	0		0	3	0	-3	0
0	0	-1	0	0		0	1	0	-1	0
G3						G4				

Figure 2.18 – Four directional high-pass filters for calculating the weighted average of luminance changes in four directions: 1: vertical, 2: diagonal (upper-left to lower-right), 3: horizontal, 4: diagonal (upper-right to lower-left)

In order to evaluate the perceptible visual quality, Chou and Li ([187]) have proposed a new metric, integrating human visual characteristics, called peak signal-to-perceptible-noise ratio (PSPNR). PSPNR only takes account of the perceptible noise as follows:

$$PSPNR = 10\ \log_{10}\left(\frac{255^2}{MSE_{JND}}\right) \tag{2.31}$$

where

$$MSE_{JND} = \frac{1}{HW}\sum_{x=0}^{H-1}\sum_{y=0}^{W-1}\left(e(x,y)_{JND}\right)^2 \tag{2.32}$$

$$e(x,y)_{JND} = \left(\left|p(x,y) - \hat{p}(x,y)\right| - JND(x,y)\right)\cdot\delta(x,y) \tag{2.33}$$

$$\delta(x,y)=\begin{cases}1,\ if\ \left|p(x,y)-\hat{p}(x,y)\right|>JND(x,y),\\ 0,if\ \left|p(x,y)-\hat{p}(x,y)\right|\le JND(x,y).\end{cases} \tag{2.34}$$

PSNR is a particular case of PSPNR ($JND(x,y)=0$ then PSPNR equals PSNR). Only errors higher than the JND profile are taken into account for calculating PSPNR. Yang et al. ([191]) extended Chou and Li's model to account for multiple channels and the combined effect of contrast masking and luminance adaptation. Thus, the spatial JND threshold for a pixel can be expressed as follows

$$\begin{aligned}JND_\theta^S(x,y)=&T^l(x,y)+T_\theta^t(x,y)\\&-C_\theta^{lt}\cdot\min\left\{T^l(x,y),T_\theta^t(x,y)\right\}\end{aligned} \tag{2.35}$$

where $T^l(x,y)$ and $T_\theta^t(x,y)$ represents the background luminance adaptation and texture masking, respectively, at a colour channel Θ, and C_θ^{tl} reflects the overlapping effect in masking. Regarding texture masking Yang et al. ([191]) propose to incorporate in $T_\theta^t(x,y)$ the edge information as follows:

$$T_\theta^t(x,y)=\beta_\theta mg_\theta(x,y)W_\theta(x,y) \tag{2.36}$$

where β_θ is an empirical parameter for each colour channel ([191]), and $W_\theta(x,y)$ is an edge-related weight of the pixel at (x,y), and its corresponding matrix W_θ is computed by edge detection followed by a Gaussian low-pass filter:

$$W_\theta=E_\theta\times h \tag{2.37}$$

where E_θ is the edge matrix of the original video frame for each colour component, detected by a Canny detector ([192]) with the sensitivity threshold of 0.5, 0.175 and 0.175 for Y, Cb and Cr respectively, and h is a k×k Gaussian low-pass filter (kernel size=7; σ. = 0.8) ([191]). The temporal effect on JND can be incorporated as the scaled amplitude of the spatial JND ([188]). JND can thus be calculated as

$$JND_\theta(x,y,t)=f\left(ild_\theta(x,y,t)\right)JND_\theta^S(x,y) \tag{2.38}$$

where $f\left(ild_\theta(x,y,t)\right)$ represents the average inter-frame luminance difference between the $(t)th$ and $(t-1)th$ frame [188].

$$ild_{\theta}(x,y,t) = \frac{p(x,y,t) - p(x,y,t-1) + bg(x,y,t) - bg(x,y,t-1)}{2} \tag{2.39}$$

where bg is the average intensity and $f_3(\ )$ is the empirical function defined in [188].

Figure 2.19 provides three examples of the estimation of local JND thresholds, for each pixel, in the first image of the encoded sequences of Akiyo, Foreman and Football (these sequences will be described in more detail in Chapter 5). The difference between decoded and original images are computed and transformed into JND units. Figure 2.19 displays the maximum weighted average of luminance difference, $mg(x,y)$, the average background luminance, $bg(x,y)$, and the JND profile according to the (2.23). The resulting picture provides a good description of image quality. The coding algorithm used was H.264/AVC (CBR=256kbps, IntraPeriod=4, NoBFrames, JM rate control [169],[170], more information in next Chapters). Black pixels denote parts of the image where distortion is at JND level 1 or less. Thus, areas of the image that have errors in the magnitude of one JND unit display small or no visible degradation, while regions of the image with higher amplitudes correlate well with the regions where coding degradation is visible.

Figure 2.19 – JND Maps (from left to right: mg, bg and JND profile; from top to bottom: Akiyo sequence, Foreman sequence, Football sequence)

Thus, bright areas correspond to particularly visible distortions predicted with the JND profile, whereas dark areas represent the predicted invisible distortions. The JND Profile is valuable since it illustrates not only the magnitude, but also the location of noticeable differences.

2.2.6 *Structural Approach*

HVS-based metrics present various problems that have been well documented in the literature ([64],[67],[112],[163],[164]). One of the most difficult problems is the limited understanding of HVS, and the true meaning of the term visual quality ([67]). The continuous increase of knowledge regarding the way HVS works, should support the improvement of better HVS-based metrics. Meanwhile, research concerning the non-HVS-based algorithm has evolved. A new scheme for a class of quality metrics, known as structural similarity (SSIM), has been proposed to implicitly model perception by taking into account the fact that HVS is adapted for extracting structural information (relative spatial covariance) from images ([193],[194]). SSIM is an objective evaluation metric that attributes perceptual degradations to structural distortions. The SSIM index is an effective measurement of perceptual global degradations in natural images ([193]). It is also a comparable metric to conventional error-based perceptual quality metrics. In essence, the SSIM index is a measurement of deviations in luminance l, contrast c, and structure s between the reference and the distorted image. Figure 2.20 show the block diagram of SSIM.

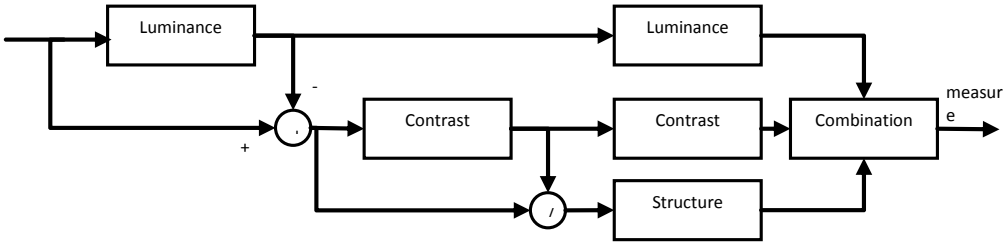


Figure 2.20 – Diagram of image similarity measurement system ([164])

Let $x = \{x_i \,|\, i = 1, ..., N\}$ and $y = \{y_i \,|\, i = 1, ..., N\}$ denote vectors from corresponding image patches in the reference and test images X and Y, correspondingly. From each patch, luminance, contrast and structural degradations are defined, respectively, by the equations $l(x, y)$ $c(x, y)$, $s(x, y)$:

$$l(x,y) = \left(\frac{2\mu_x \mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \right), c(x,y) = \left(\frac{2\sigma_x \sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \right), s(x,y) = \left(\frac{\sigma_{xy} + C_3}{\sigma_x \sigma_y + C_3} \right) \tag{2.40}$$

with μ_x and μ_y the average of each x and y, σ_{xy} the covariance of x and y, σ_x^2 and σ_y^2 the variance of each x and y, C_1, C_2 and C_3 constants to avoid instability when $\mu_x^2 + \mu_y^2$ is very close to zero ([164]). The SSIM measure is a combination of these three distortion components between signals x and y as shown below:

$$SSIM(x,y) = \left[l(x,y)\right]^{\alpha} \cdot \left[c(x,y)\right]^{\beta} \cdot \left[s(x,y)\right]^{\gamma} \tag{2.41}$$

where α, β and γ are positive parameters and are used to regulate the relative importance of each of the components. Typically, the values of these parameters are set to one. The SSIM index is computed as follows:

$$\begin{aligned} SSIM(x,y) &= \left(\frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1}\right) \cdot \left(\frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2}\right) \cdot \left(\frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3}\right) \\ SSIM(x,y) &= \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2)(\sigma_x^2 + \sigma_y^2 + C_2)} \end{aligned} \tag{2.42}$$

with $C_1 = (k_1 L)^2$, $C_2 = (k_2 L)^2$, $C_3 = C_2/2$, L corresponds to the dynamic range of the pixels (usually $2^{bits\ per\ pixel} - 1$), $k_1 = 0.01$, and $k_2 = 0.03$ ([164]). The images can be examined as a whole or on a sliding windows basis. In [195], a 11 × 11 circular-symmetric Gaussian weighting function $w = \{w_i \,|\, i = 1, ..., N\}$ was adopted with a standard deviation of 1.5 samples, normalised to sum to unity $\left(\sum_{i=1}^{N} w_i = 1\right)$. In [196], it is referred a 7 × 7 circular-symmetric Gaussian weighting function with standard deviation of 1.5 samples. If weighting is used, then the estimates of μ_x, σ_x, and σ_{xy} are adapted as follows:

$$\begin{aligned} \mu_x &= \sum_{i=1}^{N} w_i x_i \\ \sigma_x^2 &= \sum_{i=1}^{N} w_i \left(x_i - \mu_x\right)^2 \\ \sigma_{xy} &= \sum_{i=1}^{N} w_i \left(x_i - \mu_x\right)\left(y_i - \mu_y\right) \end{aligned} \tag{2.43}$$

Non-uniform weighting can be required for certain applications ([193]). For example, if some previous knowledge about the relevance of different regions in the image is accessible, then this information can be used as a weighting function ([197]). In the final step, the quality map is converted into a single quality index for the whole image. One solution is by computing the mean value of local SSIM values or using other pooling methods (fixation-based pooling [198],

percentile pooling [199], information-content-based pooling [200], perceptually pooling [201]). Local values are combined into the frame-level quality index as follows ([202]):

$$Q_i = \frac{\sum_{j=1}^{R_S} W_{ij} \cdot SSIM_{ij}}{\sum_{j=1}^{R_S} W_{ij}} \tag{2.44}$$

where R_S is the number of sampling windows used, i the i[th] video frame, j the j[th] sampling window, w_{ij} the weighting value used for the i[th] frame, and $SSIM_{ij}$ the summation of the weighted SSIM values calculated for the signal components in the j[th] sampling window in the i[th] frame. The overall quality measure for the video sequence is then calculated as defined:

$$Q = \frac{\sum_{i=1}^{F} W_i \cdot Q_i}{\sum_{i=1}^{F} W_i} \tag{2.45}$$

where F is the total number of frames, w_i the weighting value used for the i[th] frame, and Q_i the quality index calculated for the i[th] frame. Thus, for example, it is possible to allocate a smaller weight to blurred images caused by camera movement as global blur does seem to be perceived as less disturbing than the local blur ([202]).

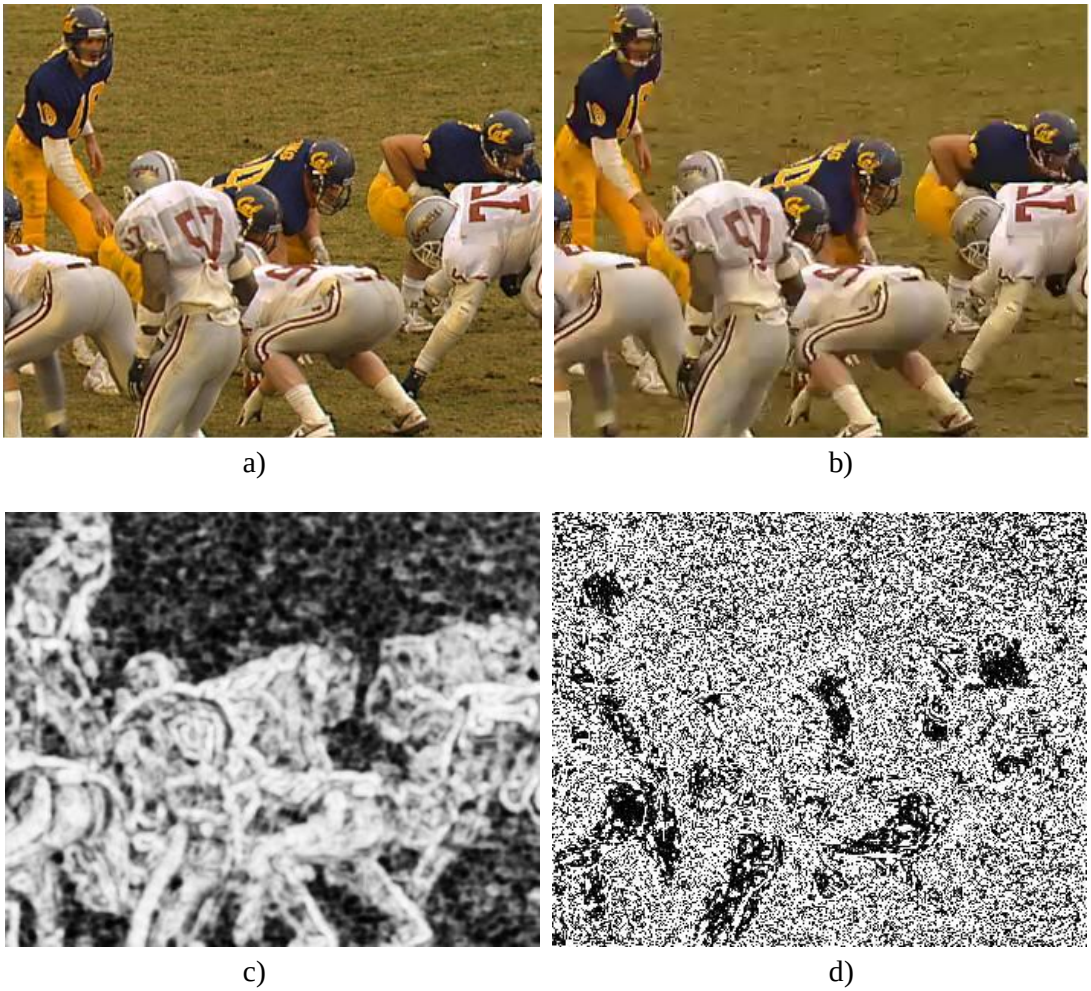
SSIM was extended to the complex wavelet domain, known as the complex wavelet structural similarity index (CW-SSIM) ([195],[203]). CW-SSIM is given by:

$$S\left(c_x, c_y\right) = \frac{2\left|\sum_{i=1}^{N} c_x, c_{y,i}^{*}\right| + K}{\sum_{i=1}^{N}\left|c_{x,i}\right|^2 + \sum_{i=1}^{N}\left|c_{y,i}\right|^2 + K} \tag{2.46}$$

where $c_x = \left\{c_{x,i} \middle| i = 1, ..., N\right\}$ and $c_y = \left\{c_{y,i} \middle| i = 1, ..., N\right\}$ are two sets of complex wavelet coefficients in the same position in the reference and test images. Spatial domain SSIM algorithm is highly sensitive to translation, scaling, and rotation of images ([204]). In CW-SSIM, because structural similarity is measured in the complex wavelet domain, it achieves high performance and a degree of translation-invariance ([67]). CW-SSIM has been used in face recognition ([205]). The SSIM has also been extended to multiple scales ([206]). However, the multi-scale SSIM does not always yield better results than its single-scale version ([207]). More extensions of SSIM have been proposed such as a gradient based approach ([208]). In [209] an image coding quality assessment is proposed based on fuzzy integral. An image is divided into

different parts (edges, textures and flat regions), after applying SSIM on the images. A weighted sum of the SSIM scores, from each of the regions, is combined to generate a content-based metric value for the image. Kandadai have extended the SSIM concept to audio structure similarity ([210]).

Figure 2.21 displays the SSIM index map of the first image of the Football video sequence. The coding algorithm used was H.264/AVC (CBR=256kbps, IntraPeriod=4, NoBFrames, JM rate control ([169],[170]), more information in Chapter 3 and Chapter 4). The absolute error map is included for comparison. Both distortion maps have been adapted to be more comprehensible. Thus, brighter areas always indicate better quality regarding a given quality/distortion measure. It can be seen that the compressed image shows variable quality across space. Nevertheless, there is no direct association between the distortion in the absolute error map and the underlying image structures (Figure 2.21 d). By contrast, the SSIM index map, Figure 2.21 c), gives perceptually uniform prediction.



a) b)

c) d)

Figure 2.21 – Football Distorted Image and its quality/distortion maps (a) original image; (b) H.264/AVC compressed image; (c) SSIM index map; (d) absolute error map

The characteristic artefacts of MPEG compression are visible in Figure 2.21 c). Typically, data compression algorithms generate smoother areas of detailed image structures. For example, the texture information of the grass of the football stadium is reduced. The SSIM index map predicts this phenomenon very well, but the absolute error map fails to predict it. The SSIM index map demonstrates the efficiency of SSIM in identifying the loss of quality in the image, by successfully predicting image quality variations across space. The Absolute Error Map does not represent the distortion existing in the different regions of the image.

2.3 Summary

This chapter discusses the prime ideas concerning video quality and its assessment. Following an introduction and background on quality assessment, the concept of quality metrics is introduced. There are two distinct classes of methods available to perform video quality assessment ([67],[68]): subjective and objective measurement's methods.

Subjective quality assessment seeks to capture, based on ratings given by human observers, the users' opinions regarding video quality. It is a particular reliable way of evaluating video quality. As a result, it is also the most efficient method to test the performance of human vision models and objective quality assessment metrics. ITU has formalised several methodologies for the subjective assessment of the visual quality of television pictures, in ITU-R Rec. BT.500 ([70],[71]), and multimedia systems in ITU-T Rec. P.910 ([72]). In this chapter, the following ITU methods were briefly described: DSIS, DSCQS, SS, SDSCE, and SSCQE. The SAMVIQ, specified by EBU, was also addressed ([81]). The methods differ, in general, in the use of different rating scales and in displaying or not a reference.

In the objective method, a metric is based only on mathematical methods without human intervention. The goal is to design quality metrics that can predict perceived video quality automatically. To deal with different requirements, distinctive objective quality assessment models have been developed, and five models have been described herein: media-layer model, parametric packet-layer model, parametric planning model, bitstream layer model, and hybrid model.

Evaluation of perceived video quality is a hard task due to the complexity of HVS. A general overview of the main characteristics of Visual Perception is provided in this Chapter. Increased knowledge in these areas fostered the development of better Vision Models and Video Quality Assessment Metrics based on HVS.

A summary of the different methods to classify objective quality metrics was presented giving particular emphasis to their main characteristics. One of the methods is the amount of

information on the reference video that is available ([60],[77],[158],[159]). In this case, objective video quality metrics can be categorised in three classes: Full Reference (FR), Reduced Reference (RR) and No Reference (NR) ([108],[130],[131],[211],[212]). If both the reference and the impaired videos are totally available, the objective metrics are classified as FR. If the reference is not available, the metric is classified as NR. If some characteristics of the reference and the impaired video are available, the objective metrics is referred as RR metric.

Various organisations stimulate the development and standardisation of the technology needed for evaluating the performance of digital video processing and communication systems. In particular, VQEG was established with the goal of studying objective video quality metrics and to provide input to ITU-T Study Groups 9 and 12 and ITU-R Study Group 6 ([78],[79]). VQEG promoted independent validation tests to assess different models. Most of the models proposed for objective quality assessment were FR. It was found that FR methods outperform RR and NR methods. Although NR models present several advantages in the industry's perspective, more research of NR methods is still required to reach the same level of prediction accuracy as the FR and RR methods ([212]). Traditional FR metrics, such as MSE and PSNR, are based solely on the differences between frames. It is a simple and fast way to predict video quality. Two different FR approaches for quality assessment, based on JND and on structural similarity (SSIM), have been presented regarding their concept, implementation and meaning. In contrast with traditional metrics, they used mechanisms to incorporate HVS or the perceptual effects of video degradation. As a result, they allow a more refined prediction of the level of degradation that a signal can suffer until a human observer notices it ([212]).

The work presented in this dissertation has been planned considering the use of subjective and objective quality assessment metrics. The subjective tests have been performed in order to gather the opinion of human observers regarding the quality of a number of representative (in terms of spatial features, motion and coding artefacts) compressed video sequences. One compression standard, the H.264/AVC, has been considered. Different objective metrics were also used, including SSIM, PSNR and PSPNR. The next chapter will introduce video coding standards.

Chapter 3. Digital Video Coding Standards Overview

In the last twenty years, several video coding standards have been developed to guarantee an efficient usage of visual information along the entire chain that covers the production, distribution, and reception of video material. Standardising allows independent implementations and guarantees that those implementations will be interoperable. This effort was mostly motivated by the need to decrease the vast quantities of data generated by a video source due to constraints in the capacity of broadcast channels and in storage space. Every standard describes the bitstream syntax and decoder semantics. Therefore, manufacturers of video hardware have the stability needed to invest in developing their products. Standardising the decoder bitstream allows the interoperability of products developed by different manufacturers. Content producers have the guarantee that their audio-visual material works with any product, and that it is not necessary to produce and manage numerous copies to satisfy the requirements of different manufacturers. As a result, it drove innovation around software and hardware encoding solutions. This strategy has played a key role in the global spread of digital video communication. Standards offer comprehensive guidance to guarantee content delivery and interoperability.

Internationally, two working groups lead the video coding standardisation processes, specifically, the ITU-T Video Coding Experts Group (VCEG) and the ISO/IEC Moving Picture Experts Group (MPEG). VCEG has targeted low bit rate video coding and their corresponding applications. Typically, for this type of application, there is a demand for high compression rates and error resilience tools. MPEG focus on higher bit rates for entertainment-quality broadcasting applications. Both institutions have generated well-known standards.

VCEG continued the work of the "Specialists Group on Coding for Visual Telephony" of the ITU-T (previously named CCITT). VCEG is associated with the standardisation process of the "H.26x" family of video coding standards. CCITT was responsible for the publication, in 1989, of the H.120 Recommendation ([213]) and, in 1993, the H.261 Recommendation ([214]). Both Recommendations were designed for video conference services. Aiming at higher compression ratios, the ITU-T started activities in 1993 with the objective of issuing a recommendation for the video coding of narrow telecommunication channels. After finalizing the H.263 Standard for

video telephony ([215]), the ITU-T VCEG started work on two further development areas: a “near-term” effort to add extra features to H.263 (resulting in Version 2 of the standard) and a “long-term effort" to develop a new standard for low bit rate visual communications. This last approach led to the draft “H.26L” standard, offering significantly better video compression efficiency than the previous ITU-T standards.

The ISO/IEC is responsible for standards such as the ISO/IEC 10918 standards (JPEG) for still picture compression, published in 1984 ([216]). In 1988, ISO/IEC established the Moving Picture Experts Group (MPEG) working group with the goal of the “development of international standards for compression, decompression, processing, and coded representation of moving pictures, audio, and their combination, in order to satisfy a wide variety of applications” ([217]). The MPEG standard families are generic standards. This means that each standard is independent of any particular application and of delivery media. However, it does not ignore the requirements of the possible applications ([218]). These generic standards are extremely important as they allow the development of VLSI and several of the basic blocks required for a large number of applications. They are the result of joint development efforts of audio and video compression experts considering the requirements of all applications. The toolbox approach, bounded by the one feature, the one tool principle, is one of the reasons for the success of the MPEG standards family.

In the first phase of MPEG standardisation, called "divergence phase," the requirements for specific applications or fields of applications are identified. Independent laboratories, universities and commercial companies, present their algorithms to fulfil specified requirements and the different approaches are compared ([219],[220]). The second phase, "convergence phase," the process continues with the selection of the most relevant coding techniques from among the techniques presented in the first phase. With the joint effort of all MPEG participants, the selected coding techniques are further refined and optimised. In the last phase, "verification" or "validation," software simulations/hardware tests are used to validate the results obtained in the previous phases.

In 1993, ISO and IEC published the MPEG-1 standard (ISO/IEC 11172) ([217],[218],[221]). The MPEG-1 standard was planned for storage applications of audio-visual information on compact disc. Its successor, MPEG-2 standard (ISO/IEC 13818), was designed for the generic coding of audio-visual information. MPEG-2 is used in DVD and for digital broadcasting ([2]). The MPEG-4 (ISO/IEC 11496) started activities in 1993, aiming not only for higher compression ratios but also the incorporation of multimedia functionalities. MPEG-4 supports additional tools for communication, access and manipulation of digital audio-visual data ([222]). In 2001, the ISO/IEC MPEG group recognised the potential benefits of H.26L and

created the Joint Video Team (JVT), including experts from MPEG and VCEG ([223]). JVT's main task was to develop the draft H.26L "model" into a full international standard. In fact, the outcome is two identical standards: ISO MPEG-4 Part 10 of MPEG-4 and ITU-T H.264/AVC. The "official" title of the standard is Advanced Video Coding (AVC); however, it is widely known by its ITU document number, H.264/AVC ([6]).

While MPEG-1, MPEG-2, and MPEG-4 were aimed at digital content coding, following MPEG standards evolved towards content description (MPEG-7) ([224]) and transactions (MPEG-21) ([225]). MPEG-1 and MPEG-2 provide interoperable ways of representing audiovisual content, commonly used on digital storage media and broadcast media. MPEG-4 extends this to many more application areas through features like its extended bit rate range, scalability, error resilience, seamless integration of different types of 'objects' in the same scene, interfaces to digital rights management systems and powerful ways to build interactivity into content. MPEG-7 aims to provide tools for describing all forms of multimedia content delivered by the broadest possible range of networks and terminals ([224]). MPEG-7 has descriptive elements that range from very 'low-level' signal features like colours, shapes and sound characteristics, to high-level structural information about content archives. However, automatic feature extraction is not within the scope of the MPEG-7 project. The focus is on standard description elements of multimedia data designated by Descriptor and Description Scheme (DS). A language was standardised to specify Description Schemes and Descriptors in MPEG-7. It is referred to as Description Definition Language (DDL) ([224]). A hierarchical structure of Descriptors and Description Schemes is used to describe the multimedia data. With MPEG-7, it is possible to exchange information about multimedia content in interoperable ways, making it easier to find content and identify just what end-users wanted to find. MPEG-7 information can be added by broadcasts or personal video recorders so that managing multimedia content in large content repositories can be facilitated ([224]). It is important to guarantee interoperability when there is a solution for protecting the existing digital assets. Digital Rights Management (DRM) is becoming a necessity to protect the value of content and to guarantee services' viability. Existing systems are using non-standardised protection mechanisms. The MPEG-21 Multimedia Framework initiative aims to enable the transparent and augmented use of multimedia resources across a wide range of networks and devices ([225],[226]). MPEG-21 is based on two essential concepts: the definition of a fundamental unit of distribution and transaction (the Digital Item) and the concept of Users interacting with Digital Items. The Digital Items can be considered the "what" of the Multimedia Framework (e.g. a video archive, a music album) and the Users can be considered the "who" of the Multimedia Framework ([225]). The goal of MPEG-21 can thus be understood as the defining of the technology needed to provide support to users in order to

access, exchange, consume and manipulate Digital Items in an efficient, transparent and interoperable way ([225]). The MPEG-21 multimedia framework identifies the key elements needed to support the multimedia delivery chain and record the relationships between, and the operations supported by them ([225]).

To confirm whether the standard satisfies the identified requirements it has to be evaluated and checked against the identified requirements in the same way a product is tested against the product specifications. Two operational tools are used for verification: the working model and core experiments. For MPEG-1 and MPEG-2, the verification tests consisted of formal subjective tests aimed at evaluating the quality of either audio or video signals processed using the MPEG algorithms. In MPEG-4, new types of tests were undertaken using optimised assessment methods ([225]).

This Chapter provides a brief introduction to some of the most well-known international video coding standards (MPEG-1 [221], MPEG-2 [2], MPEG-4 [227], and H.264/AVC [6]). These standards allow the efficient broadcast and storage of different video sources in a diverse environment. They are based on a hybrid motion-compensated predictive video coding approach obtaining high compression ratios by removing both spatial and temporal redundancies existing in video sources.

3.1 The MPEG 1 Video Standard

The MPEG-1 video coding algorithm, although flexible, was optimised to give its best performance at bit rates around 1.2 Mbps working with a picture spatial resolution of 350 pixels per 250 lines and frame repetition rates between 24 and 30 images per second ([218],[221]). As it is a generic standard, MPEG-1 was developed in response to the need for a common format for representing compressed video on different Digital Storage Media (DSM) such as CDs, DATs, Winchester disks and optical drives ([218]). Applications using compressed video on DSM need to be able to perform a number of operations in addition to normal forward playback of the video sequence. In order to meet the needs of these application's several features were identified ([218],[221]):

- Random Access.
- Fast Search (Forward/Reverse).
- Reverse Playback.
- Error Robustness.
- Coding/decoding delay.

Two key techniques are used to encode video data: spatial and temporal techniques. The first technique is accomplished by removing redundant information within a picture. The second technique involves exploiting the similarities between successive frames of a video signal by only encoding the difference. The process by which the amount of movement contained between pictures is computed is traditionally known as motion estimation prediction. The data gathered from this method is then used in the motion compensating prediction. The MPEG-1 video algorithm can produce three types of encoded pictures, by using the following three different coding modes ([218],[221]):

- *Intra mode* (I-pictures), pictures encoded individually without using temporal prediction (without reference to any other picture).
- *Predicted mode* (P-pictures), pictures encoded using motion compensated prediction from the previous picture as a result of containing a reference to the previous picture.
- *Bi-directional mode* (B-pictures), generating inter-coded pictures that use bi-directional temporal prediction.

B-frames may be encoded using either forward prediction where reference is made to a picture in the past, backward prediction where reference is made to a future picture, or to an image in the past and an image in the future.

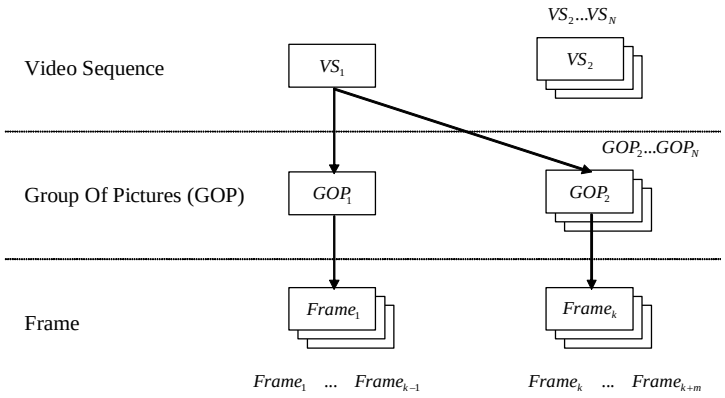


Figure 3.1 – Hierarchical structure of the MPEG-1 video bitstream

The encoded data is organised into a layered structure that permits the integration of the different coding modes. This structure comprises six hierarchical layers: Video Sequence, Group of Pictures (GOP), Picture, Slice, Macroblock and Block. The first three layers are displayed in Figure 3.1. One sequence contains one or more Group of Pictures (GOP) ([218],[221]). A GOP is composed of one or more encoded pictures. It is used as a random access unit. The GOP length is the period (frequently expressed in frames) at which an Intra-

frame occurs. The GOP length may be dictated by the random access requirements. A picture is divided into slices, macroblocks and blocks. A block is a picture area of eight pixels by eight lines either of the luminance or the chrominance components (used as a DCT unit). A macroblock (MB) associates four blocks of luminance with the spatially corresponding block of each chrominance component. Motion compensation uses macroblocks as its unit. A slice is a collection of an integer number of macroblocks, in raster-scan order. Usually, a slice is a horizontal stripe within a frame. The first slice of a picture must start with the upper-left macro block of that picture, and the last slice must end with the lower-right macro block.

The use of temporal interpolation implies a rearrangement of the order of the coded pictures before broadcasting. I and P pictures must be transmitted before interpolated (B) frames. A typical GOP in display order might be as in (3.1) whereas the bitstream order would be as in (3.2).

$$I_0 B_1 B_2 P_3 B_4 B_5 P_6 B_7 B_8 P_9 B_{10} B_{11} I_{12} \tag{3.1}$$

$$I_0 P_3 B_1 B_2 P_6 B_4 B_5 P_9 B_7 B_8 I_{12} B_{10} B_{11} \tag{3.2}$$

The MPEG-1 Video compression algorithm is a hybrid of motion compensation and the transform coding algorithm. Figure 3.2 (a) shows the major components of a MPEG-1 video encoder. Video source data is encoded on a block-by-block basis. The first picture is usually encoded without motion compensation. The DCT transforms a block by converting the spatial domain pixels into transform domain coefficients. The quantisation block uses a quantiser to diminish the number of levels of the transformed coefficients. The difference between the original block and the predicted block, called the prediction error, is then encoded.

The quantised transform coefficients are de-quantised and the inverse DCT is applied producing a decoded picture. This picture can be used as a reference for the encoder and the decoder. Thus, the following pictures can use motion compensation to reduce temporal redundancies. It is possible to find, using motion estimation techniques, a motion vector that gives a best match for a source picture block between the source picture and a reference picture from the picture store. The motion compensation uses the motion vector and generates a predicted error block from the reference picture.

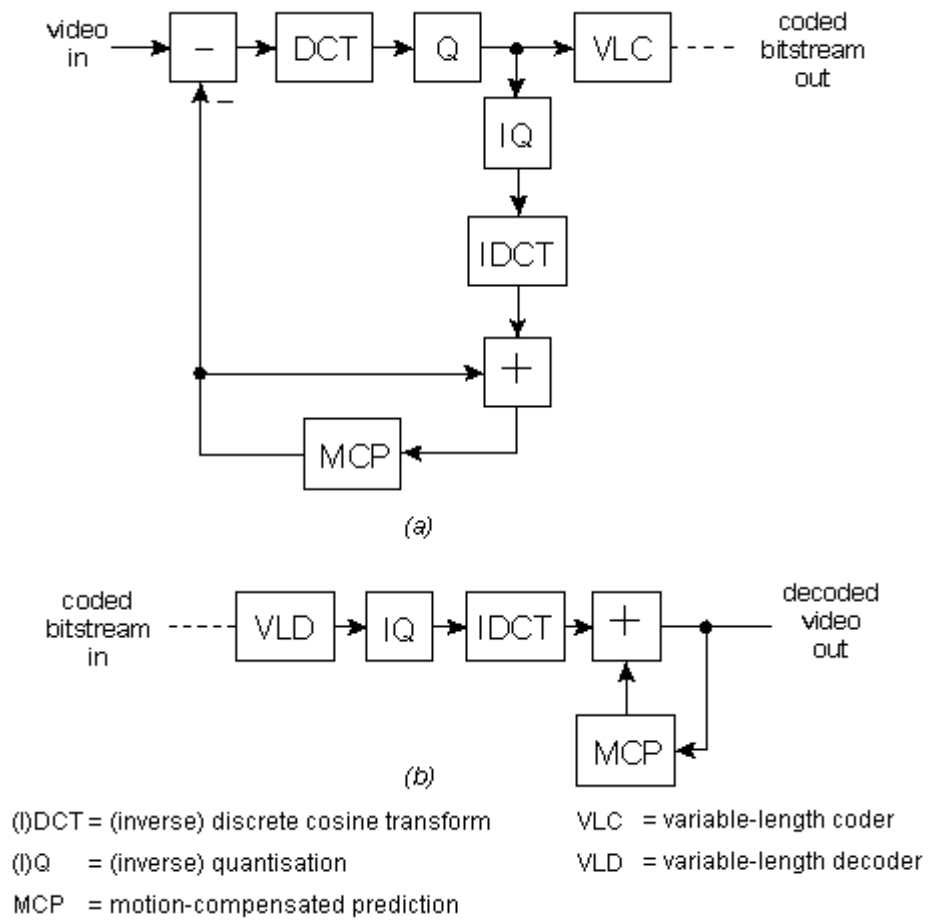


Figure 3.2 – (a) Motion-compensated DCT coder; (b) motion compensated DCT decoder

In an MPEG-1 Video Decoder (Figure 3.2 (b)), the first step of the inverse process is to reconstruct the quantised DCT coefficients. The reconstructed coefficients are subsequently added to the prediction error.

3.2 The MPEG 2 Video Standard

In 1991, MPEG started a second phase of work with the goal of developing a standard to cover a wider range of applications rather than just storage and retrieval in DSM, offering much higher picture resolutions and bit rates (MPEG-2) ([2]). From the early stages, the main application of MPEG-2 was the all-digital transmission of broadcast TV quality video at coded bit rates between four and nine Mbps (Mega bits per second). However, the MPEG-2 syntax was adapted to be suitable to other applications such as those at higher bit rates and sample rates. The most significant enhancement over MPEG-1 is the addition of syntax for the efficient coding of interlaced video (e.g. 16x8 block size motion compensation, Dual Prime, et al). Several other more subtle improvements (e.g. 10-bit DCT DC precision, non-linear quantisation, VLC tables, and improved mismatch control) are included, which provide a clear improvement to coding efficiency, even progressive video. Other key features of MPEG-2 are

the scalable extensions which permit the division of a continuous video signal into two or more coded bitstreams representing the video at different resolutions, picture quality (i.e. SNR), or picture rates.

3.2.1 *Scalability*

The MPEG-2 toolkit includes tools for 'scalable' coding. As mentioned before, useful video can be reconstructed from parts of the full bitstream. The entire bitstream is organised in layers, beginning with a base layer (that can be decoded by itself) and added enhancement layers to reduce quantisation distortion or improve resolution. Scalability is also a useful tool for error resilience on prioritised media ([228]). The drawback of scalability is that some coding efficiency is lost because of the extra overhead.

MPEG-2 video has several scalable modes, which include spatial scalability, temporal scalability, SNR (signal-to-noise ratio) scalability, and data partitioning. The MPEG-2 support incorporates combinations of these basic scalability tools.

Spatial scalability allows multi-resolution coding. A single video source is divided into a base layer (lower spatial resolution) and an enhancement layer (higher spatial resolution). Figure 3.3 displays the encoding and decoding process of an MPEG-2 Video Encoder that supports spatial scalability. For example, a CCIR 601 video can be down-sampled to SIF format using spatial decimator. The SIF bitstream can be encoded independently from the enhancement layer and thus serve as the base layer video. The enhancement layer bitstream corresponds to additional spatial information. The prediction image in the enhancement layer encoder is the weighted sum of the temporal prediction image of the enhancement encoder and the spatial prediction image, up sampled from the base layer encoder (Figure 3.3 - WT – Weighter; STA – Spatio-Temporal Analyser). Weights may be adapted on a MB level. Spatial scalability is a suitable tool for applications where the interworking of video standards is required as well as for simultaneous broadcast. The drawback is that there exists some bit rate penalty due to the overhead and there is a moderate increase in complexity.

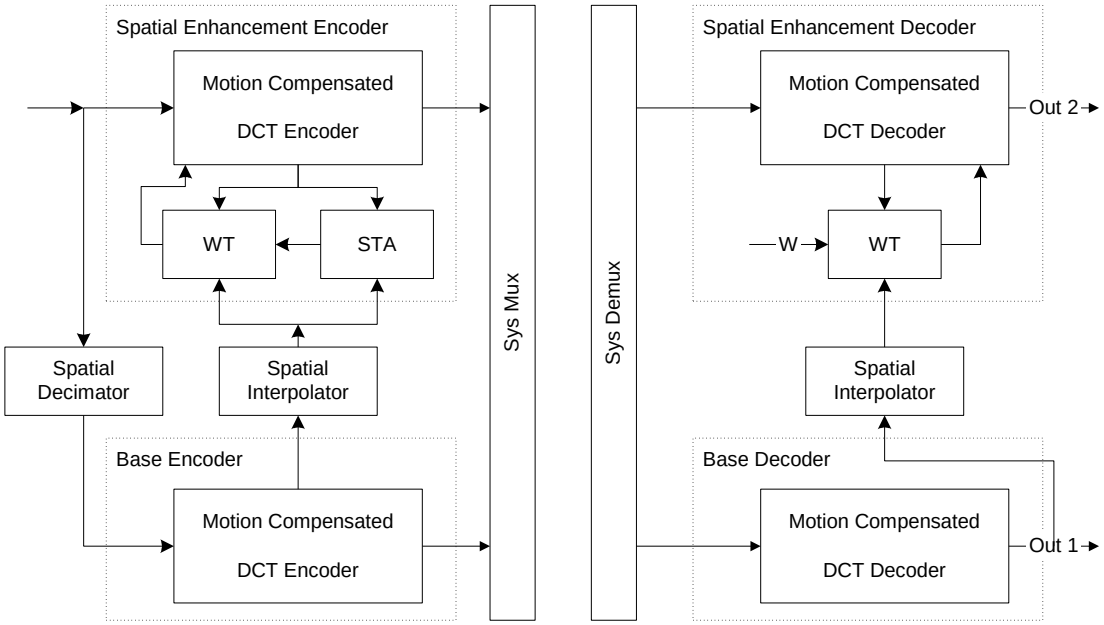


Figure 3.3 – Block diagram for MPEG-2 codec with spatial scalability

The SNR scalability provides a mechanism for transmitting two-layer service with the same spatial resolution but different quality levels. Figure 3.4 illustrate the encoding and decoding process with SNR scalability.

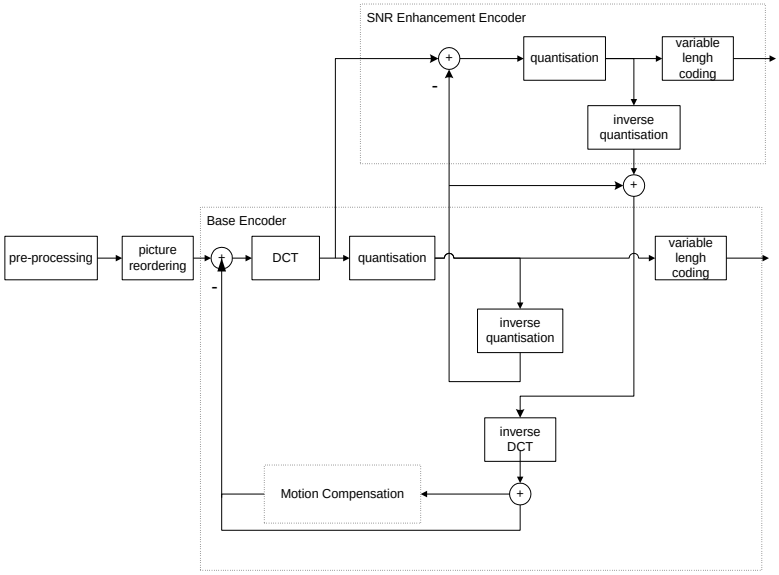


Figure 3.4 – Block diagram for MPEG-2 codec with SNR scalability

The encoding of the lower layer uses a coarse quantisation step size typically for low-capacity channels. The enhancement layer corresponds to the difference between the original and the coarse-quantised signals. It is encoded with a finer quantiser to generate an enhancement bitstream for high-quality video applications. SNR scalability can be used in scenarios where SDTV/HDTV or in video services with multiple qualities.

In temporal scalability, the base layer corresponds to encode at a lower frame rate, and the intermediate frames encoded in a second bitstream using the first bitstream reconstruction as a prediction (Figure 3.5).

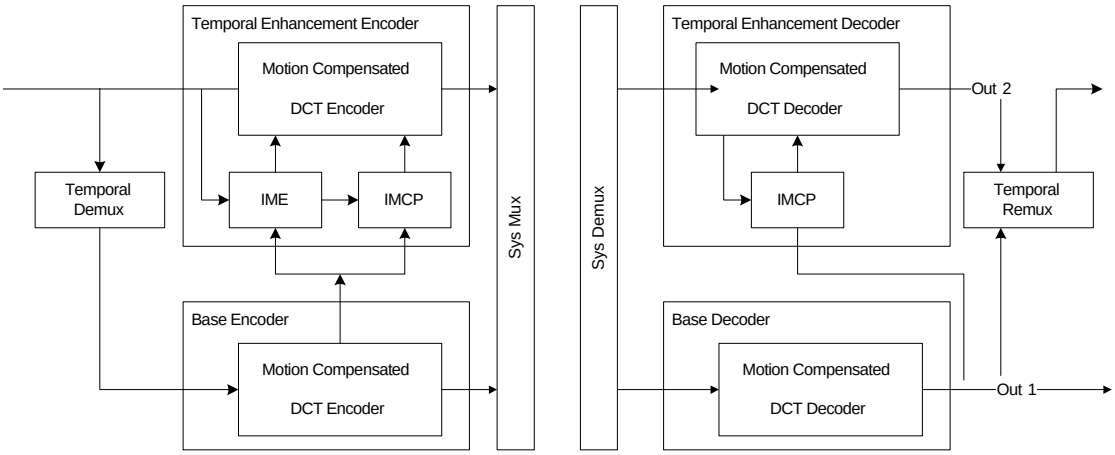


Figure 3.5 – Block diagram for MPEG-2 codec with temporal scalability

The Temporal Demux sends alternating frames to the base encoder and the enhancement encoder. The Temporal Enhancement Encoder uses a motion estimation/compensation technique (Interlayer Motion Estimator - IME; Interlayer Motion Compensated Predictor - IMCP).

Data partitioning is a frequency-domain method that breaks the block of 64 quantised transform coefficients into two bitstreams (Figure 3.6).

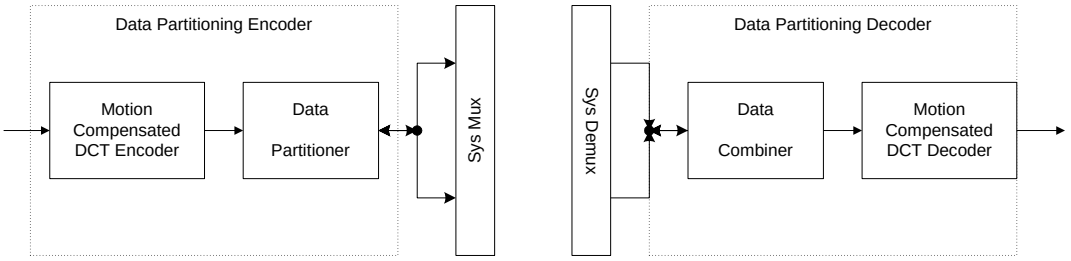


Figure 3.6 – Block diagram for MPEG-2 codec with data partitioning scalability

The first, higher-priority bitstream contains the more critical lower frequency coefficients and side information (such as headers, DC values, motion vectors). The second, inferior priority bitstream carries higher-frequency AC data. It is appropriate when two transmission channels are available. Unlike the other scalable tools, both layers are needed to reconstruct the original

3.2.2 *MPEG-2 Profiles and Levels*

MPEG-2 is a generic standard. Different algorithmic elements or 'tools, developed for distinct uses, are combined into a single syntax to meet the requirements of various applications ([229]).

However, to prevent all codecs from having to support the implementation of the full syntax, the concept of "Profiles" was introduced. A Profile is a defined subset of the entire bitstream syntax ([2]). MPEG-2 specifications define two non-scalable profiles, Simple Profile (SP) and Main Profile (MP), and three scalable profiles: SNR Profile, Spatial Profile and High Profile (HP) ([2]). The profile's definition process guarantees that a higher profile is a superset of a lower one. It is still possible to observe a large variation in the performance of encoders/decoders depending on the values taken by parameters in the bitstream, within the limits imposed by the syntax of a given profile. A solution was defined within each Profile with the concept of levels. A level is a defined set of constraints imposed on parameters in the bitstream (Low Level, Main Level, High 1440 Level, and High Level).

Profile		Simple (I, P) 4:2:0	Main (I, P,B) 4:2:0	SNR (I, P,B) 4:2:0	Spatial (I, P,B) 4:2:0	High (I, P,B) 4:2:2; 4:2:0	Multiview (I, P,B) 4:2:0	4:2:2 (I, P,B) 4:2:2; 4:2:0
High Level	Pels/line		1920			1920	1920	1920
	Lines/frame		1152			1152	1152	1152
	Frame/s		60			60	60	80
	Mbps		80			100	130	300
High-1440 Level	Pels/line		1440		1440	1440	1440	
	Lines/frame		1152		1152	1152	1152	
	Frame/s		60		60	60	60	
	Mbps		60		60	80	100	
Main Level	Pels/line	720	720	720		720	720	720
	Lines/frame	576	576	576		576	576	576
	Frame/s	30	30	30		30	30	30
	Mbps	15	15	15		20	25	50
Low Level	Pels/line		352	352			352	
	Lines/frame		288	288			288	
	Frame/s		30	30			30	
	Mbps		4	4			8	

Table 3.1 – Profiles and levels in MPEG-2

Not all profiles support all the levels. Table 3.1 indicates the allowable picture types (I, P, B), pels/line and lines/picture, picture format and maximum bit rate (for all layers in case of scalable bitstreams), for each profile/level pair. For example, SP uses no backward or interpolated prediction. Therefore, in this case, no picture reordering is required. This makes SP suitable for low-delay applications such as video telephone or video conferencing. MP adds support for B pictures. Thus, picture quality may be increased and a higher degree of compression may be achieved. The MP decoder should also decode MPEG-1 video bitstreams. The SNR profile adds support for two levels of picture quality. The addition of an extra quantisation stage does not essentially change its nature and the codec works like an MP codec.

The error introduced by the first quantisation is itself quantised, run-length and VLC coded, and transmitted as the enhancement layer. Two additional profiles, developed after the final approval of MPEG-2, are included in Table 3.1. The first, the 4:2:2 profile, is capable of working with pictures that have a colour resolution of 4:2:2 and a higher bit rate. The 4:2:2 profile was approved in January 1996. The other is the Multiview Profile (MVP). It is possible using existing MPEG-2 video coding tools to efficiently encode two video sequences issued from two cameras shooting the same scene with a small angle between them.

3.3 The MPEG 4 Video Standard

The development of MPEG-4 reflects new trends in the standardisation of multimedia information resulting from the merging of three worlds: telecommunications, audio-visual, and computing ([222],[230]). Initially, MPEG defined MPEG-4 as a standard centred on Very Low Bit Rates or on Very High Compression Efficiency. The main goal was to considerably improve the video compression efficiency of the existing hybrid DCT-based coding schemes. At the same time, the ITU-T Low Bit rate Coding (LBC) group started producing the first results on the near-term hybrid coding solution (the ITU-T H.263 standard [231]). In July 1994, the Grimstadt MPEG meeting marked an important modification in the direction of MPEG-4. Firstly, it conducted an in-depth analysis of trends in the audio-visual world based on the convergence of the TV/film/entertainment, computing, and telecommunications worlds. After this analysis, a new goal was established. The MPEG group redefined MPEG-4 as an "emerging coding standard that supports news ways (notably content-based) for communications, access, and manipulation of digital audio-visual data" ([222],[232]). Thus, MPEG-4 became the first standard that understands an audio-visual scene as a composition of objects (audio, visual, or audio-visual), in accordance with a script that describes their spatial and temporal relationship ([233]). Users have the possibility of interacting with the audio-visual content of a scene and mixing synthetic and natural audio and video information. The MPEG-4 Proposal Package Description (PPD) describes eight new or improved functionalities clustered in three main groups ([233],[234]):

- *content-based interactivity*, addressing the ability to interact with meaningful objects in an audio-visual scene (four key functionalities: content-based multimedia data access tools, content-based manipulation and bitstream editing, hybrid natural and synthetic data coding, and improved temporal random access),
- *high compression efficiency*, important not only to enable low bit rate applications, but also to provide the ability to efficiently code multiple views of

a scene (two key functionalities: improving coding efficiency and coding of multiple concurrent data streams),

- *universal accessibility*, meaning that access to audio-visual data should be available for a wide range of storage and transmission media (two key functionalities: robustness in error-prone environments and content-based scalability).

The MPEG-4 provides functionalities included in existing standards such as MPEG-1 or MPEG-2. In addition, some of the features referred to above, namely content-based interactivity, not only represent additional features but also a major evolution in the way that audio-visual information has been represented. Structures, like the region or the object that represents the visual information in MPEG-4 are more complex than the pixel. These structures must be easily associated with meaningful semantic units. The next section will describe some of the most important coding innovations.

3.3.1 *MPEG-4 Parts*

MPEG-4 provides a standardised way to represent the content of audio-visual objects, describe the composition of these objects in compound media scenes, encode, multiplex and synchronise the data associated with the media objects, transport the media presentation over different channels, and interact with the audio-visual scene at the receiver's end. The multi-part ISO/IEC 14496 series, "Information technology – Coding of audio-visual objects," defines the set of technologies for compression, encoding and delivery of complex audio-visual scenes composed of distinctive media objects: video objects (natural or synthetic video), audio objects, still images, text and vector graphics, computer-animated images ([235]). Each part covers a limited segment of the whole specification. The most known parts are MPEG-4 part 2 and the MPEG-4 part 10 (MPEG-4 AVC/H.264). The MPEG-4 Standard consists, at the time of writing, of 28 parts, under the general title "Information technology — Coding of audio-visual objects", and briefly described in the next paragraphs.

Part 1: Systems.

The System specification describes synchronization and multiplexing of video and audio ([236],[237]).

Part 2: Visual.

The Visual specification contains definitions of the bitstream syntax, bitstream semantics and the related decoding process. It does not specify the encoders that can be optimised in different implementations ([238],[239]).

Part 3: Audio.

The Audio specification integrates many different types of audio coding: natural sound with synthetic sound, low bit rate delivery with high-quality delivery, speech with music, complex soundtracks with simple ones, and traditional content with interactive and virtual-reality content (e.g. Advanced Audio Coding (AAC), Audio Lossless Coding (ALS), Scalable Lossless Coding (SLS), Structured Audio, Text-To-Speech Interface (TTSI), and others) ([240],[241]).

Part 4: Conformance testing.

The Conformance part describes procedures to verify whether bitstreams and decoders meet requirements specified in other parts (ISO/IEC 14496 (parts 1, 2 and 3) and for ISO/IEC 14496-6:2000) ([242]).

Part 5: Reference software.

The Reference software specification provides software implementations of the ISO/IEC 14496 (parts 1, 2, 3, and 6) including normative and non-normative tools ([243]).

Part 6: Delivery Multimedia Integration Framework (DMIF).

DMIF stipulates a session protocol for the management of multimedia streaming over generic delivery technologies ([244]).

Part 7: Optimised reference software for coding of audio-visual objects.

It specifies the encoding tools that improve both the execution and the quality for the coding of visual objects as defined in ISO/IEC 14496-2 ([245]).

Part 8: Carriage of ISO/IEC 14496 contents over IP networks.

It specifies a framework for the carriage of MPEG-4 contents over IP networks ([246]). It also provides guidelines to design payload format specifications for the detailed mapping of MPEG-4 content into several IP-based protocols.

Part 9: Reference hardware description.

It provides hardware designs of the principal video tools for demonstrating how to implement the other parts of the standard ([247]).

Part 10: Advanced Video Coding (AVC).

It provides a compression format for video signals ([6]). It is technically identical to the ITU-T H.264 standard. The ISO/IEC 14496-10:2010 standard includes the description of advanced video coding (AVC) and associated extensions to support scalable video coding (SVC) and multiview video coding (MVC).

Part 11: Scene description and application engine.

It specifies a coded representation of interactive audio-visual scenes and applications ([248]).

Part 12: ISO base media file format.

It provides a file format for storing time-based media content ([249],[250]). It is a general format forming the basis for a number of other more specific file formats.

Part 13: Intellectual Property Management and Protection (IPMP) Extensions.

It provides a common Intellectual Property Management and Protection (IPMP) processing, syntax and semantics for carry IPMP tools in the bitstream ([251]).

Part 14: MP4 file format.

It specifies the MP4 file format that defines the storage of MPEG-4 content in files ([252],[253]).

Part 15: Advanced Video Coding (AVC) file format.

It provides a file format for the compressed video streams using any of the coding tools defined in MPEG-4 part 10 (e.g. AVC, SVC or MVC) ([254],[255]).

Part 16: Animation Framework eXtension (AFX).

It describes MPEG-4 Animation Framework eXtension (AFX) model for representing 3D Graphics content ([256],[257]).

Part 17: Streaming text format..

It provides a method for the coding of text at very low bit rate as one of the multimedia components within an audiovisual presentation ([258]).

Part 18: Font compression and streaming.

It specifies font data representation, compression and streaming, providing an efficient mechanism to embed font data in MPEG-4 encoded presentations ([259]). It also defines MPEG-4 Text profiles and levels.

Part 19: Synthesised texture stream.

It specifies the broadcast of synthesized texture data ([260]).

Part 20: Lightweight Application Scene Representation (LASeR) and Simple Aggregation Format (SAF).

This part defines a scene description format (LASeR) and an aggregation format (SAF) respectively adequate for describing and producing rich-media services to resource-constrained devices such as mobile phones ([261],[262]).

Part 21: MPEG-J Graphics Framework eXtensions (GFX).

It specifies a lightweight programmatic environment for advanced interactive multi-media applications oriented for limited resource's devices such as mobile phones ([263]).

Part 22: Open Font Format.

It defines the Open Font Format (OFF) specification, the TrueType™ and the Compact Font Format (CFF) outline formats, and the TrueType hinting language ([264],[265]).

Part 23: Symbolic Music Representation (SMR).

It specifies Symbolic Music Representation (SMR) ([266]).

Part 24: Audio and systems interaction.

It specifies the desired joint behaviour of MPEG-4 File Format and MPEG-4 Audio codecs ([267]).

Part 25: 3D Graphics Compression Model.

It describes a model for connecting 3D Graphics compression tools defined in MPEG-4 to graphics primitives defined in any other standard, specification or recommendation ([268]).

Part 26: Audio Conformance.

It describes how tests can be designed to verify whether compressed data and decoders meet requirements specified by MPEG-4 Audio ([269],[270]).

Part 27: 3D Graphics conformance.

3D Graphics Conformance summarises the requirements, and defines how conformance can be tested ([271],[272]). Guidelines are given on creating tests to verify decoder conformance.

Part 28: Composite font representation.

It specifies the Composite Font Representation – an XML-based document format that allows combining individual component font resources into a single virtual font ([273]). This part, at the time of writing, is under development.

3.3.2 *MPEG-4 Visual Coding Innovations*

An important goal of MPEG-4 was the integration of different types of media objects. An efficient coded representation of these objects was a pre-requisite, so each type of object in the scene has its own optimal representation. This is in contrast with, for example, MPEG-2 Video where all objects are merged together, transformed into pixels, and coded using hybrid DCT coding. Consider the example of Figure 3.7 where a MPEG-2 encoder is receiving rendering images in real time. The graphics engine would compute the appearance of any virtual objects, from the selected viewpoint. If the viewpoint or one of the objects were to move, each video frame would be different, and the MPEG-2 encoder would use its coding tools to encode the image differences.

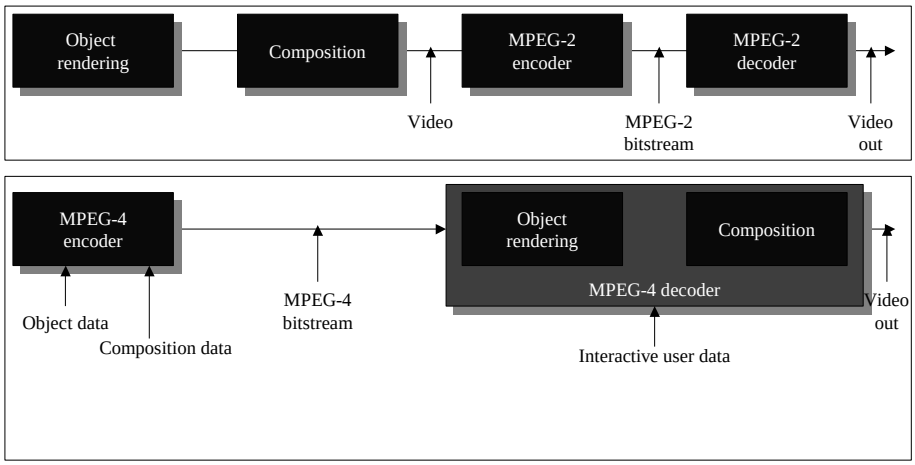


Figure 3.7 – MPEG-4 versus MPEG-2 encoding process

An MPEG-4 encoder can directly handle graphic instructions. Thus, after the object is decoded the scene is composed according to the scene description. MPEG-4 defines efficient coded representations for the following types of objects ([233],[274],[275],[276]):

- Natural audio (including special speech codecs);
- Synthetic sound;
- A text-to-speech interface;
- Arbitrarily shaped video and stills;

- Facial and body animation;
- Generic 2-D and 3-D "computer generated objects."

By using shape in the coding process, it is possible to achieve better subjective quality, increased coding efficiency as well as an object-based video representation [277]. MPEG-4 visual allows the transmission of arbitrarily shaped video objects (VO's) ([233],[274],[275],[276]). Figure 3.8 shows the complete hierarchy of an MPEG-4 video bitstream.

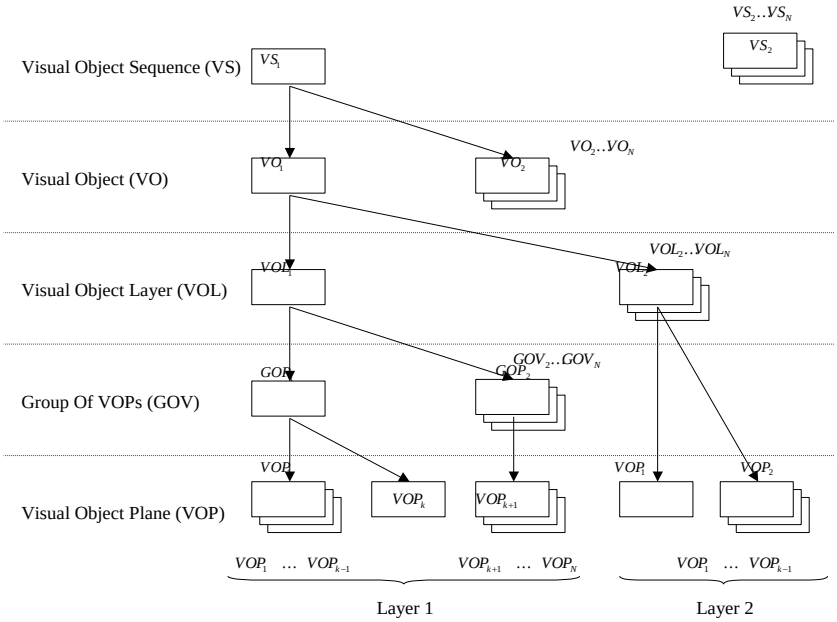


Figure 3.8 – Hierarchical structure of the MPEG-4 video bitstream

A MPEG-4 bitstream contains one or more visual sequences (VS) ([275]). The input video source can be segmented into a number of arbitrarily shaped image regions (video objects planes - VOP) ([232]). This process of segmentation is outside the scope of the MPEG-4 standard. Usually, video is segmented according to its use. The shape and location of each VOP can vary from frame to frame. Consecutive VOP's belonging to the same region in a scene are referred to as video objects (VO's) ([233],[274],[275],[276]). The shape, motion, and texture information of VOP's belonging to the same VO are encoded and transmitted, or encoded into a separate video object layer (VOL). If the VO is scalable, it may be divided and sent in two or more video object layers (VOL). One of these VOL's is the base layer and the remaining VOL's are the enhancement layers. Each layer corresponds to a certain spatial resolution or image quality. Furthermore, a set of successive VOP's can be clustered in order to form a group of VOPs (GOV). The GOV carries header information, which is useful for random access and resynchronisation.

For coding purposes, each VOP (rectangular size or not) is encoded using a block-based hybrid DPCM/transform coding technique ([275]). A VOP is a snapshot in time of a video object layer and is composed of luminance and chrominance components plus shape information. The encoder contains two parts: shape coding and the traditional motion and texture coding applied to the same VOP. Shape information is represented by *alpha masks*. An alpha mask defines the level of transparency of a VOP, corresponding to zero value for completely transparent pixels. Texture coding as well as motion estimation and compensation are similar to MPEG-2 encoding techniques. For an arbitrary shaped VOP, shape is first encoded and then texture information is partitioned into no overlapping macroblocks. MPEG-4 extends the types of macroblocks supporting different types: transparent (macroblocks completely inside the VOP), opaque (macroblocks completely outside the VOP) and boundary (macroblocks at the boundary of the VOP).

Although the usage of shape and thus the possibility of content-based functionalities is a major innovation, it is not the only one. For example, the frame-based parts of the video coding algorithms are improved regarding efficiency and error resilience. ,Some of the most important new tools for video encoding are now listed ([227],[233],[274],[275],[276]):

- A wider selection of block sizes and the flexibility to dynamically select motion compensation block size.
- Improved prediction of motion vectors to more efficiently support the representation of complicated motion information.
- Advanced prediction of AC and DC intra-coefficients so that the encoder can represent more efficiently texture data.
- Multiple reference frame selection, allowing the encoder to find the best match across multiple video frames.
- Quarter pixel motion estimation, rendering motion with a higher degree of accuracy, even for non-translational slowly moving objects. MPEG-4 H.264/AVC further improves motion rendition by using high-quality interpolation filters.
- A finer quantiser step resolution for chrominance blocks than for luminance blocks, thus reducing colour-smearing problems occasionally visible with MPEG-2.
- The MPEG-4 AVC entropy coder, that is context-adaptive and uses an arithmetic coder rather than MPEG-2's Huffman coder.

3.3.3 *MPEG-4 Visual Profiles*

MPEG-4 presents a toolbox using profiles with distinct solutions for various application settings. Like in MPEG-2, MPEG-4 defines different levels of complexity: specifications of the constraints (e.g. bit rate, sampling rate, object to memory size, etc.) that are associated with a profile. One major difference is the object-based audio-visual representation model that is supported by MPEG-4. In addition, while MPEG-2 provides profiles only for video, MPEG-4 extends to other types of media ([227]):

- Visual profiles are defined as the visual object types that can be used in the scene.
- Audio profiles do the same for audio.
- Graphics profiles are defined as the BIFS nodes that specify the graphical elements that can be composed in the scene.
- Scene graph profiles define the supported scene composition capabilities, such as 2 or 3 dimensionality, interaction capabilities, and support for e.g.: translation, rotation, scaling.
- Object descriptor profiles define the capabilities of the synchronization layer and the object descriptors tools.

Different profiles support multiple application environments. They allow manufacturers to use a subset of the rather large MPEG-4 toolbox ([227]). Profiles have been defined for two main reasons: to ensure interoperability between MPEG-4 implementations and to allow conformance to the standard to be tested. In MPEG-4, several profiles have been defined. Only some visual profiles will be introduced in this section. There are 19 visual profiles, defined for different applications and for different classes of visual object types: rectangular video, arbitrarily shaped video, still visual, and synthetic visual object types. The MPEG-4 Visual profiles for coding ‘natural’ video scenes are listed in Table 3.3 (rectangular video, arbitrarily shaped video, and still visual object types). These profiles range from the so-called Simple Profile (coding of rectangular video frames) through profiles for arbitrarily-shaped and scalable object coding to profiles for the encoding of studio-quality video ([275],[278]). Table 3.2 presents the profiles for coding ‘synthetic’ video (animated meshes or face/body models) and the hybrid profile (incorporates features from synthetic and natural video coding) ([278]).

MPEG-4 Visual Profile	Main features
Basic Animated Texture	2D mesh coding with still texture
Simple Face Animation	Animated human face models
Simple Face and Body Animation	Animated face and body models
Hybrid	Combines features of Simple, Core, Basic Animated Texture and Simple Face Animation profiles

Table 3.2 – MPEG-4 Visual profiles for coding synthetic or hybrid video ([278])

MPEG-4 Visual profile	Main features	Video Object Types
Simple	Low-complexity coding of rectangular video frames.	Rectangular Video
Advanced Simple	Coding rectangular frames with improved efficiency and support for interlaced video.	Rectangular Video
Advanced Real-Time Simple (ARTS)	Coding rectangular frames for real-time streaming.	Rectangular Video
Core	Basic coding of arbitrarily-shaped video objects.	Arbitrarily Shaped Video
Main	Feature-rich coding of video objects.	Arbitrarily Shaped Video
Advanced Coding Efficiency	Highly efficient coding of video objects.	Arbitrarily Shaped Video
N-Bit	Coding of video objects with sample resolutions other than 8 bits.	Arbitrarily Shaped Video
Simple Scalable	Scalable coding of rectangular video frames.	Rectangular Video
Fine Granular Scalability	Advanced scalable coding of rectangular frames.	Rectangular Video
Core Scalable	Scalable coding of video objects.	Arbitrarily Shaped Video
Scalable Texture	Scalable still texture with improved efficiency and object-based features.	Still Visual
Advanced Scalable Texture	Scalable still texture with improved efficiency and object-based features.	Still Visual
Simple Studio	Object-based coding of high-quality video sequences.	Arbitrarily Shaped Video
Core Studio	Object-based coding of high-quality video with improved compression efficiency.	Arbitrarily Shaped Video

Table 3.3 – MPEG-4 Visual profiles for coding natural video ([278])

The simple profile and advanced simple profile have been considerably used by the industry for mobile applications and streaming on networks. The simple profile has been used to encode rectangular video with intra (I) and predicted (P) VOPs. The simple profile allows the use of three compression levels with bit rates from 64 kbits/s in level 1 to 384 kbits/s in level 3. The advanced simple profile is also used to encode rectangular video with intra (I) and predicted (P) VOPs. However, it is enhanced to add bidirectional (B) VOPs for better coding efficiency. Furthermore, it supports six compression levels (0 –5). The bit rate from the first four levels ranges from 128 to 768 kbits/s. Interlaced coding is associated with levels 4 and 5, with bit rates from 3 to 8 Mbits/s.

Profile \ Object types	Simple	Advanced Simple	Advanced Real-Time Simple	Core	Main	Advanced Coding Efficiency	N-bit	Simple Scalable	Fine Granular Scalability	Core Scalable	Scalable Texture	Advanced Scalable Texture	Simple Studio	Core Studio	Simple Face Animation	Simple Face and Body Animation	Basic Animated Texture	Animated 2D Mesh
Simple	✓																	
Advanced Simple	✓	✓																
Advanced Real-Time Simple	✓		✓															
Core	✓			✓														
Advanced Core	✓			✓								✓						
Main	✓			✓	✓						✓							
Advanced Coding Efficiency	✓			✓		✓												
N-bit	✓			✓			✓											
Simple Scalable	✓							✓										
Fine Granular Scalability	✓	✓							✓									
Core Scalable	✓			✓				✓		✓								
Scalable Texture											✓							
Advanced Scalable Texture												✓						
Simple Studio													✓					
Core Studio													✓	✓				
Basic Animated Texture											✓				✓		✓	
Simple Face Animation															✓			
Simple FBA																✓		
Hybrid	✓			✓							✓				✓		✓	✓

Table 3.4 – MPEG-4 Visual profiles and objects ([278])

Table 3.4 lists each of the MPEG-4 Visual profiles (left-hand column) and visual object types (top row) ([278]). The table entries specify which object types are contained within each profile. For example, a codec compatible with Advanced Simple Profile must be capable of coding and decoding Simple and Advanced Simple objects.

3.4 The H.264/AVC Video Standard

The H.264/AVC Standard is the outcome of a joint research initiative of the ITU-T VCEG and the ISO/IEC MPEG standardisation committee ([6]). It builds on the same concepts of previous standards such as MPEG-1, MPEG-2, MPEG-4 part 2, H.261, H.263. VCEG started the standardisation project in 1998, originally called H.26L. The aim was to improve coding efficiency in comparison with existing video coding standards for an extensive range of applications. A Call for Proposals was issued in 1998 by VCEG, and a first draft was available in 1999. In 2001, VCEG and MPEG formed a JVT with the goal of developing a standard and the reference software. The reference software was named joint model (JM). In 2003, the first phase of the standard was completed and submitted for formal approval. ISO/IEC adopted the standard under the name of MPEG-4 Part 10 AVC, and ITU-T adopted the standard with the name of H.264 ([6]). An overview of the first phase of H.264/AVC is presented in ([8]), and a

comprehensive introduction, including all the extensions and amendments, can be found in ([6],[7],[8],[9],[278],[279],[280],[281],[282],[283],[284],[285],[286],[287]). The first phase of H.264/AVC ([6]) was essentially aimed at supporting entertainment-quality video. Professional applications have more requirements such as the support of higher video resolutions. Thus, an extension of the joint project was initiated to include new extensions to the H.264/AVC standard and support. In 2005, these extensions were concluded and designated as fidelity range extensions (FRExt) ([6]). In 2007, the number of the high profiles increased to eight profiles ([6]). More information on H.264/AVC FRExt can be found at [284] and [288] for the 2005 and 2007 versions, respectively. In recent years, scalable video coding (SVC) ([285],[289],[290]), and multiview video coding (MVC) ([6],[291],[292],[293]) were developed as two amendments to H.264/AVC. It is still unsure if this effort to standardise these extensions will be successful in practical and professional applications. For example, scalable bitstreams have the advantage of allowing straightforward adaptation, but they experience a loss in rate-distortion performance when evaluated with single-layer coding ([294]).

Section	What it describes
0. Introduction	It contains a brief summary of the goals and main features of the standard.
1–5. Scope, References, Definitions, Abbreviations, Conventions	How H.264 fits with other published standards; terminology and conventions used in the document.
6. Data formats and relationships	It defines the essential formats assumed for video and coded data; main ways of deriving relationships between coded units.
7. Syntax and semantics	It describes the syntax or bitstream formats, in tables, and provides an explanation regarding the meaning and allowed values of each syntax element (semantics).
8. Decoding process	It describes, in detail, all the stages of processing required to decode a video sequence from H.264 syntax elements.
9. Parsing process	It explains the processes required to extract syntax elements from a coded H.264 bitstream.
A. Profiles and levels	Profiles define subsets of video coding tools; levels define limits of decoder capabilities.
B. Byte stream format	The syntax and semantics of a stream of coded NAL units.
C. Hypothetical reference decoder	A hypothetical decoder ‘model’ that is used to determine performance limits.
D. Supplemental enhancement information	Information that may be associated in an H.264 bitstream that is not essential for decoding.
E. Video usability information	Information about display of the coded video that is not essential.
G. Scalable Video Coding	A self-contained extension to the H.264/AVC standard that supports Scalable Video Coding (SVC).

Table 3.5 – Overview of the H.264 standard document ([279])

The most recent version of Recommendation H.264 ([6]) is a document of over 550 Pages. It includes normative content; that is, important instructions that must be conformed by H.264 codecs. It also contains informative content. At the time of writing this document, it is organised as shown in Table 3.5. Note that as amendments or updated versions of the standard are

published, further sections or annexes may be added ([278],[279]). H.264/AVC has achieved important progress in coding efficiency due to its different encoding modes such as flexible block size motion estimation, quarter pixel motion compensation, spatial intra prediction, approximate DCT, variable block size partition, multiple reference frames, in-loop de-blocking filters and context-based adaptive binary arithmetic coding, etc ([279],[295]). A concise review of the H.264/AVC standard is given in this section.

3.4.1 *Technical Description of H.264/AVC Coding Tools*

The H.264/MPEG-4 design covers a Video Coding Layer (VCL), and a Network Abstraction Layer (NAL). The VCL defines the efficient representation of the video. The NAL converts the VCL representation of the video into a format suitable for transport layers or storage media ([7]). All data are contained in NAL units (NALU). Each NALU contains an integer number of bytes, where the first byte is a header, and the remaining bytes contain the payload data. The main NALU types and their corresponding payload are shown in Table 3.6. NALUs are divided into two categories, VCL and non-VCL. The first consist of the data representing the values of the samples in the video images; the second contains any associated supplementary information that improves the usability of the decoded video signal but does not influence the normative decoding process ([283]).

NAL unit type	Content of NALU	NALU type class
1	Coded slice of a non-IDR picture	VCL
2	Coded slice data partition A	VCL
3	Coded slice data partition B	VCL
4	Coded slice data partition C	VCL
5	Coded slice of an IDR picture	VCL
6	Supplemental enhancement information (SEI)	non-VCL
7	Sequence parameter set (SPS)	non-VCL
8	Picture parameter set (PPS)	non-VCL
9	Access unit delimiter	non-VCL
10	End of sequence	non-VCL
11	End of stream	non-VCL

Table 3.6 – Types of NAL units

NAL abstracts the VCL data in a generic format for use in both packet-oriented and bitstream systems, defining byte-stream and packet-based formats ([286]). In the first case, the NAL specifies the precise pattern of the start code prefix. This allows the encoded video to be delivered as an ordered stream of bytes containing start codes. Thus, circuit-switched transport layers, such as H.320, H.324M or MPEG-2, and the decoder, can identify the structure of the bitstream ([7]). In the second case, the data packets that are framed by the system transport

protocol are specified. The waste of data is avoided due to carrying the prefix for applications RTP/UDP/IP ([7]). A NAL unit is divided in non-VCL and VCL NAL units.

The non-VCL NAL unit includes extra information like parameter setting. Earlier standards contained header information about slice, picture, or sequence. This information was encoded at the beginning of the element ([286]). In H.264/AVC, two types of parameter sets can be used: sequence parameter sets (SPS) that apply to the coded video sequence, and picture parameter sets (PPS) that apply to the decoding of one or more individual pictures within a coded video sequence. If a packet containing this type of information is lost, then all associated data with the header information is useless. To prevent this problem, packets are broadcast synchronously as self-contained, in a real-time multimedia environment ([286]).

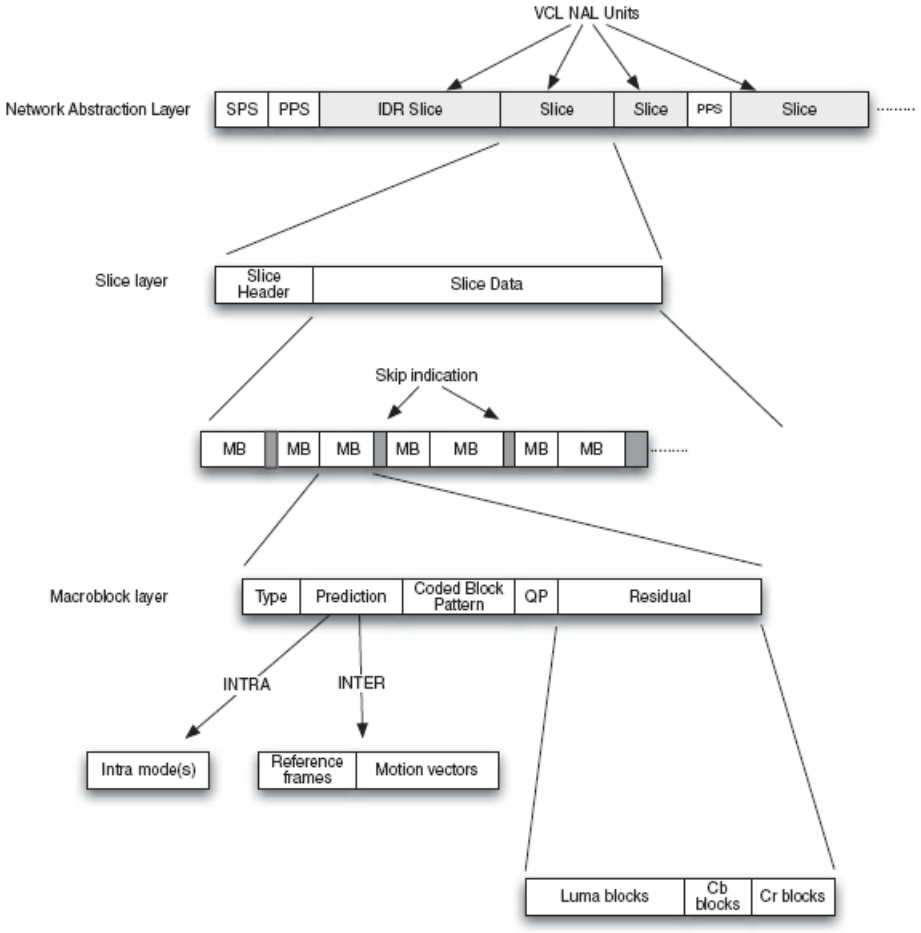


Figure 3.9 – Syntax overview ([279])

The VCL NAL unit contains the video coded data ([286]). The coding layers are the coded video sequence, picture, slice, and macroblock layers ([6]). Higher layers include lower layers. Figure 3.9 illustrates the syntax organisation of H.264/AVC, consisting of five levels of information: SPS, PPS, Slice level, MB and Block ([279],[282]). A coded video sequence starts

with an Instantaneous Decoder Refresh (IDR) Access Unit, formed by one or more IDR slices (a special type of Intra coded slice). The following video pictures are coded as slices. The video sequence finishes when the broadcast ends or a new IDR slice is received (a new coded sequence is going to start).

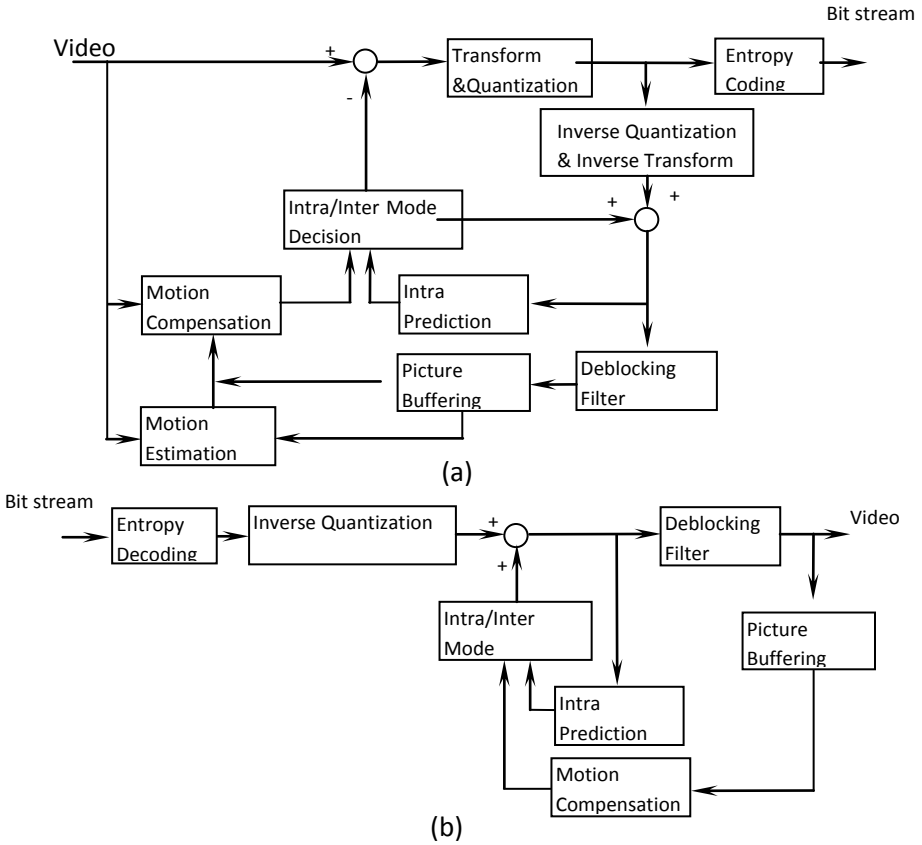


Figure 3.10 – The block diagram of H.264 Video Encoder (a) and Decoder (b) ([286])

Similar to previous coding standards, H.264 does not clearly specify a codec. Instead, it defines the syntax of an encoded video bitstream and the method of decoding this bitstream. Figure 3.10 exemplifies a typical H.264/AVC encoder and decoder. The majority of the basic functional blocks (prediction, transform, quantisation, entropy encoding) could already be found in previous standards (MPEG-1, MPEG-2, MPEG-4, H.261, H.263). The exception is the deblocking filter. Nevertheless, significant differences are found in the details of the various functional blocks.

Every picture of a video sequence is divided into fixed-size macroblocks. Each of the macroblocks covers a rectangular area of 16x16 samples of the luminance component. The number of chrominance samples depends on the chrominance sampling. For example, in the

case of 4:2:0 it corresponds to 8x8 samples of each for the two chrominance components ([285]).

In H.264/AVC, the macroblocks are processed in so called slices where a slice is typically a group of macroblocks processed into a raster scan order ([7]). Slices correspond to regions of a given picture that can be decoded independent of each other. The slice is the main concept in H.264/AVC. The information of two higher layers over slice is in the sequence parameters and picture parameters. Note that those parameters are set to be used either directly or indirectly by each slice.

Five different slice-types are supported: I-, P-, B-, SI,- and SP-slices. In I slices all macroblocks are coded without referring to any other pictures of the video sequence (Intra mode). In a P-slice, all macroblocks are predicted using a motion compensated prediction with one reference frame and in a B-slice with two reference frames. The remaining two slice types are SP (switching P) and SI (switching I) slices that are used for an efficient switching between two different bitstreams.

The H.264/AVC encoder may choose between intra and inter coding. Intra coding can generate access points to the encoded sequence. Different spatial prediction modes are available to decrease spatial redundancy in the source signal, for a single picture ([286]). Inter coding decreases temporal redundancy among different pictures using motion vectors for block-based inter prediction. Prediction is obtained from deblocking the filtered signal of preceding reconstructed pictures. The deblocking filter decreases the blocking artefacts at the block boundaries. Different motion vectors and intra prediction modes may be selected. A transform is used to reduce spatial correlation of the prediction residual before it is quantised. The prediction residual is the difference between the original input samples and the predicted samples for the block. Lastly, motion vectors or intra prediction modes are combined with the quantised transform coefficient information and encoded using an entropy code such as context-adaptive variable length codes (CAVLC) or context adaptive binary arithmetic coding (CABAC) ([286]).

The H.264/AVC standard introduces several new functionalities such as intra prediction in the spatial domain, hierarchical transform with (4x4, 8x8) integer DCT transforms and (2x2, 4x4) Hadamard transforms, multiple reference pictures in inter prediction, deblocking filter, CAVLC or CABAC entropy coding. This section will briefly present an overview of H.264/AVC coding tools. More detailed information on H.264/AVC can be found at ([6],[7],[8],[285],[289],[290], [291],[292],[293],[295],[296],[297],[298]).

3.4.2 *Intra Prediction*

One of the key features of the H.264/AVC standard is combining the transform coding with the intra-prediction in the spatial domain ([279]). Intra-coding, coding a macroblock by itself and without temporal prediction has been used in previous standards. This technique increases the quantity of encoded bits and therefore, the bit rate. For a typical block, there is quite a high correlation between samples in the block and samples that have a common border to the block. Thus, Intra prediction can reduce bit rate by using samples from contiguous, previously reconstructed (but unfiltered for deblocking) blocks to predict the values in the current block. The residual signal between the original and the predicted block is coded by using a transform coding technique. H.264/AVC supports, for luminance samples, nine intra-prediction modes for 4 × 4 blocks, four intra-prediction modes for 8 × 8 blocks, and four intra-prediction modes for 16 × 16 blocks. It is also supported by four modes for each chrominance block. The modes are selected according to the picture's characteristics. In regions of a picture that contain small details it is appropriate to use the intra 4 × 4 mode and in parts of a picture with smooth content the intra 16 × 16 mode is appropriate ([286]). Another prediction mode for the luminance component is called I-PCM. It is provided primarily for coding anomalous picture content. In this mode, pixels are sent with no prediction or transformation.

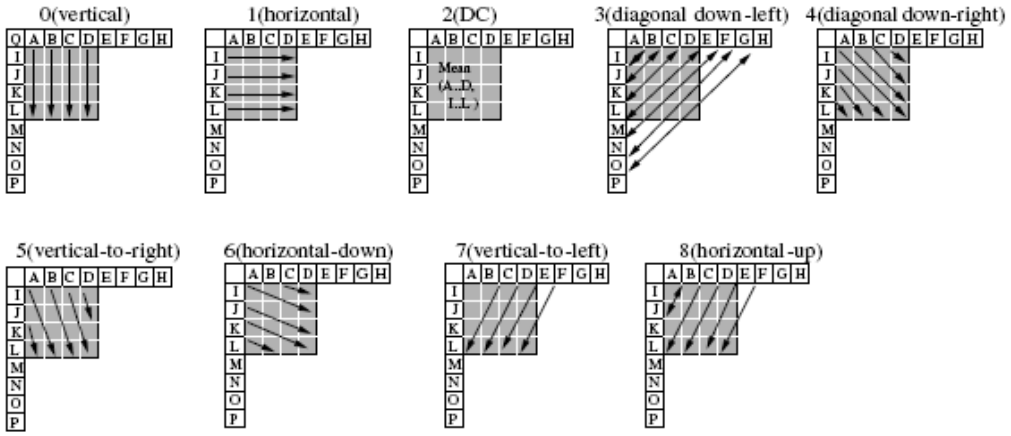


Figure 3.11 – Nine modes for 4x4 Intra Prediction ([281])

The nine modes for Intra Prediction for 4 x 4 blocks are illustrated in Figure 3.11. The shaded area is the 4x4 luminance block that is to be predicted, and above and to the left of the previously reconstructed samples [A, B, …, M] are the reference pixels in the neighbouring coded blocks. Each mode indicates prediction direction except the DC mode, in which the predicted value is the average of all the available reference pixels. Since an extensive description of the Intra prediction process in H.264/AVC is not the main topic of this dissertation, [279],[287] can be consulted for further details.

3.4.3 *Inter Prediction*

Inter prediction is the technique of predicting a block of luminance and chrominance samples from a reference picture, for exploiting the temporal redundancies that exist between successive frames. The process requires selecting a prediction region, generating a prediction block and computing the residual error of the prediction by subtracting the prediction from the original block of samples. The residual is then encoded and transmitted. In H.264/AVC, the pictures can be partitioned into macroblocks of 16 x 16 luminance samples that can be partitioned into smaller blocks sizes up to 4 x 4.

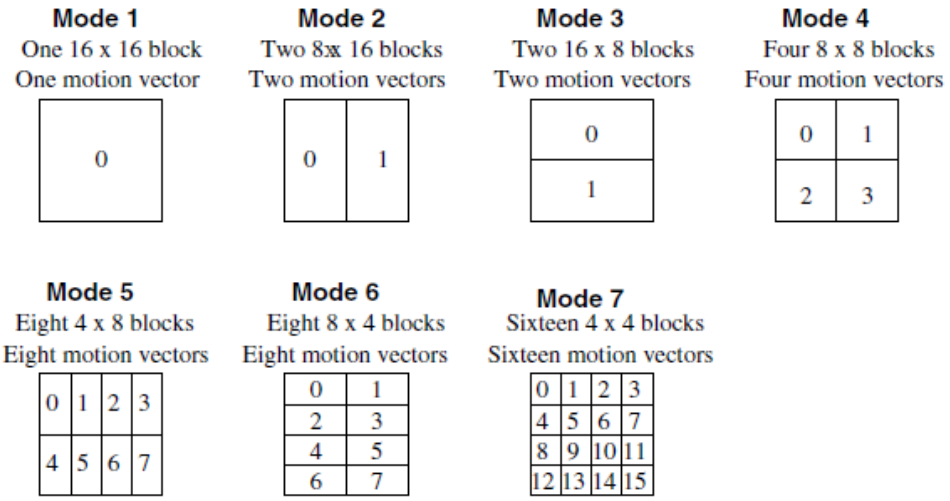


Figure 3.12 – Inter-frame prediction modes (dividing a MB into sub-blocks) ([281])

H.264/AVC allows a total of seven different block prediction sizes, including 16x16, 16x8, 8x16, 8x8, 4x8, 8x4, and 4x4 (Figure 3.12). Smaller motion compensation blocks have the potential to improve the quality of the prediction. In particular, it allows better handling of fine motion details and the motion compensated residual data can be reduced. The occurrence of blocking artefacts is avoided and the subjective video quality improved ([281]). Nevertheless, reducing the size of the block requires a larger number of bits to signal the motion vectors and extra data for the type of partition. The selection of the partition size depends on the characteristics of the input source. Normally, a big partition size is suitable for homogeneous areas of the image, and a short partition size may be more appropriate for detailed areas.

The motion compensation algorithm may be substantially enhanced by allowing motion vectors to be determined with higher spatial accuracy, compared to the previous coding standards ([281]). A motion vector corresponds to a two-dimensional vector, used for inter prediction, that provides an offset from the coordinates in the decoded picture to the coordinates in a reference picture. Each motion vector is differentially coded from the motion vectors of neighbouring blocks ([279]). The motion vector may point to integer, half- or quarter-sample positions in the

luminance component of the reference picture. Half- or quarter-sample positions are generated by interpolating the samples of the reference picture ([279]). Preceding standards have been based mainly on half-pixel accuracy, with quarter-pixel accuracy only available in H.264/AVC. The process is described in sub–section 8.4.2.2 "Fractional sample interpolation process" of the H.264/AVC standard and it will be briefly explained.

As shown in Figure 3.13, the positions referred with upper-case letters A – U, within shaded blocks represent reference picture samples at integer sample positions, and the positions labelled with lower-case letters a-o, within un-shaded blocks represent reference picture samples at fractional sample positions ([299]). The values of half-sample positions for the luminance prediction are obtained by applying a 6-tap filter, horizontally and vertically, with tap values (1, -5, 20, 20, -5, 1). The values of the luminance prediction values at quarter-sample positions are interpolated by averaging samples at integer and half-sample positions.

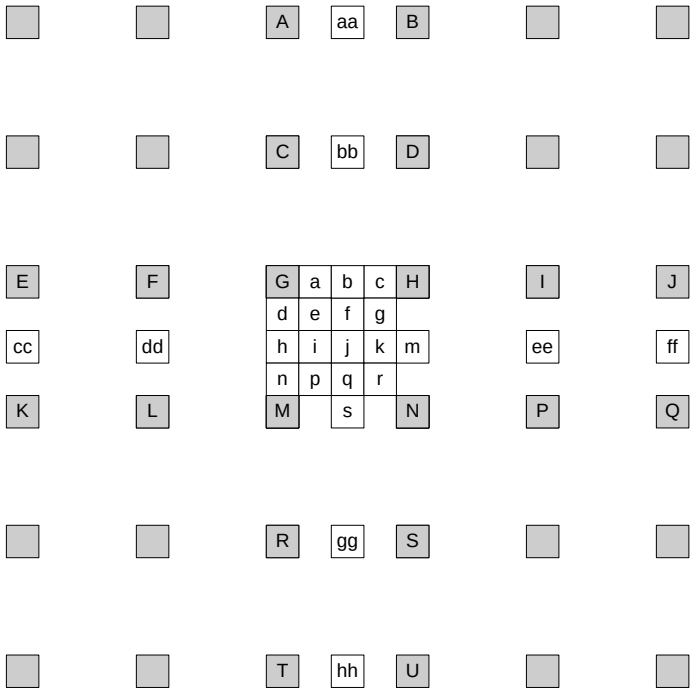


Figure 3.13 – The sub-position pixels to be interpolated and the supporting integer pixels ([299])

First, the intermediate values b1 and h1 are determined by applying the 6-tap filter in equation (3.3) and equation (3.4), respectively.

$$b1 = (E - 5F + 20G + 20H - 5I + J) \tag{3.3}$$

$$h1 = (A - 5C + 20G + 20M - 5Q + S) \tag{3.4}$$

The values of the half-pixel b and h are obtained by equation (3.5) and equation (3.6), respectively.

$$b = (b1 + 16) >> 5 \tag{3.5}$$

$$h = (h1 + 16) >> 5 \tag{3.6}$$

The value of the half-pixel position j is obtained according equation (3.7),

$$j = (cc - 5dd + 20h1 + 20ee - 5ff + gg + 512) >> 10 \tag{3.7}$$

where the intermediate values, denoted as cc, dd, ee, ff and gg, are obtained in a similar approach to h1. The values of the 1/4-pixel positions, a, c, d, f, i, k, l, and n, are computed by averaging with upward rounding of the two nearest samples at integer and half-pixel positions. For example, to determine the value of a it is computed as follows.

$$a = (G + b + 1) >> 1 \tag{3.8}$$

The values of quarter-samples positions, e, g, m, and o, are determined by averaging with upward rounding of the two nearest samples at half-pixel samples positions in the diagonal direction. The quarter-sample e is calculated by the following equation (3.9).

$$e = (b + h + 1) >> 1 \tag{3.9}$$

The sub-positions in between chrominance pixels are interpolated by bilinear filtering.

Another difference regarding previous MPEG standards, is the support of multiple reference frames. Thus, more than one previously coded picture can be used as a reference in inter-frame picture coding ([281]). Any picture type can be selected for reference. B-slices differ from P-slices as blocks can be estimated by a weighted average of two distinct MCP blocks. The Decoded Picture Buffer (DPB) is a buffer containing decoded pictures that can be used as a reference, output reordering, or output delay ([6]). Two lists of reference pictures, list 0 and list 1, are maintained at both the encoder and decoder. List 1 is only used by B-slices, while list 0 can be used by both P-slices and B-slices. These two lists contain short-term and long-term reference frames. A reference picture is short-term by default. It is identified by its frame number. Its behaviour is similar to that observed in previous standards. A picture, after being encoded, is decoded and marked as a short-term picture, meaning that is available for prediction. A long-term picture normally corresponds to older pictures that may be used for prediction. It is particularly valuable when a specific pattern is constantly used as background or in a video sequence with repeated transitions, such as interviews. The number of short-term and long-term reference frames store in DPB cannot exceed 16 pictures.

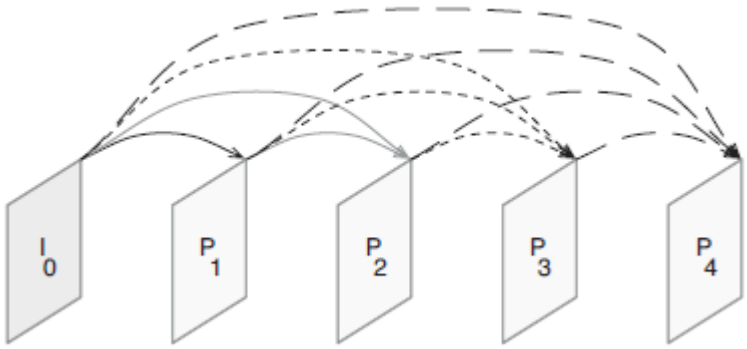


Figure 3.14 – Multiple Reference Frame Selection for Motion Compensation

Figure 3.14 ([279]) displays a prediction structure with an Intra frame and multiple P frames in which all the earlier encoded slices are accessible as reference frames. Slice P4 is predicted from slices I0, P1, P2 and P3. Slice P3 is predicted from slices I0, P1 and P2. Slice P2 is predicted from slices I0 and P1. Slice P1 is predicted from slice I0. The use of multiple reference frames might reinforce the error-resilience of the H.264 coded bitstream. Nevertheless, from an implementation perspective on both the encoder and decoder side, the use of multiple reference frames corresponds to an additional processing delay, increased implementation complexity, and higher memory requirements ([281]).

3.4.4 *Transform and Quantisation*

The Discrete Cosine Transform (DCT) has represented the typical approach for transform coding for both image and video coding in recent years ([2],[214],[215],[221]). DCT is typically performed on 8 × 8 blocks so that the energy of a transform block in the frame to be coded is compacted into a smaller number of frequency coefficients. One of the drawbacks of the 8 × 8 DCT is that it is intrinsically a floating point operation and as a result can be approximated in different ways in an integer format ([300]). Various alternatives have been proposed for replacing DCT. For example, the JPEG2000 standard uses wavelets.

In the case of H.264/AVC, three transforms are employed for different applications ([7]): 4x4 integer transform on 4x4 blocks for the luminance residual data, 2x2 transform on 2x2 of chrominance DC coefficients in any macroblock, and 4x4 Hadamard transform on the 4x4 luminance DC coefficients in intra macroblocks predicted in 16x16 mode ([280]).

The most important transform is the 4x4 integer transform. This transform is usually referred to as high correlation transform (HCT) ([301],[302]). The 4x4 HCT is applied to 4x4 predicted residual blocks of the luminance component and for all blocks of chrominance components. The forward ($\overline{H}$) and inverse ($\overline{H}_v$) transform are represented in matrix format, as follows ([280]):

$$\overline{H} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 2 & 1 & -1 & -2 \\ 1 & -1 & -1 & 1 \\ 1 & -2 & 2 & -1 \end{bmatrix} \overline{H}_v = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1/2 & -1/2 & -1 \\ 1 & -1 & -1 & 1 \\ 1/2 & -1 & 1 & -1/2 \end{bmatrix} \tag{3.10}$$

This matrix is an integer approximation of 4x4 DCT. The second transform, is a 4x4 Hadamard transform with matrix $\hat{H}$.

$$\hat{H} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & -1 & 1 & -1 \end{bmatrix} \tag{3.11}$$

This is used in addition to the first one when the macroblock is predicted in mode Intra 16x16. It transforms all 16 DC coefficients of the already transformed blocks of the luminance signal. The third and last transform is a 2x2 Hadamard transform. It transforms the 4 DC coefficients of each chrominance component. Its matrix $\hat{H}$ is shown in equation (3.12).

$$\hat{H} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \tag{3.12}$$

After transform, quantisation is applied. The H.264 uses a scalar quantiser. By default, the coefficients in a macroblock are quantised by the quantisation step. The quantisation step Q_{step}, is a function of the quantisation parameter, QP, an integer ranging from 0 to 51. The quantisation step increases two times for every increment of 6 in QP and can be represented as follows

$$Q_{step}(QP) = Q_{step}(QP\%6) \bullet 2^{floor(QP/6)} \tag{3.13}$$

H.264/AVC also provides support for perceptual-based quantisation. Besides the default quantisation matrix, an H.264/AVC encoder can specify customised quantisation matrices. These matrices are encoded at the sequence or picture level.

3.4.5 *Deblocking Filter*

In the H.264/AVC codec the deblocking filter is part of the recommendation while in MPEG-4 Visual it is an optional part of the recommendation ([278],[299]). It is applied within the motion prediction loop. As the name suggests, it is used to decrease the blockiness introduced by the discontinuity of motion and block-based transform and quantisation, usually appearing as a

high-frequency noise ([281]). As a result, better subjective and objective quality is achieved and the capacity to predict other pictures improved.

The deblocking filter is applied to the vertical and horizontal edges of 8×8 and 4×4 transform blocks. For each edge, the strength of the filter varies with the coding type of the two neighbouring blocks, the quantisation parameters, the presence of non-zero coefficients, and in the case of motion-compensated blocks, the values of their corresponding motion vectors.

Filtering one block involves these steps ([299]). The first step is to determine the value of the boundary strength (BS) between neighbouring 4x4 luminance blocks. This parameter is used to choose an appropriate filter. It varies from 0 (no filtering) to 4 (strongest filtering), and assess the potential degree to which the two neighbouring blocks suffer from blockiness.

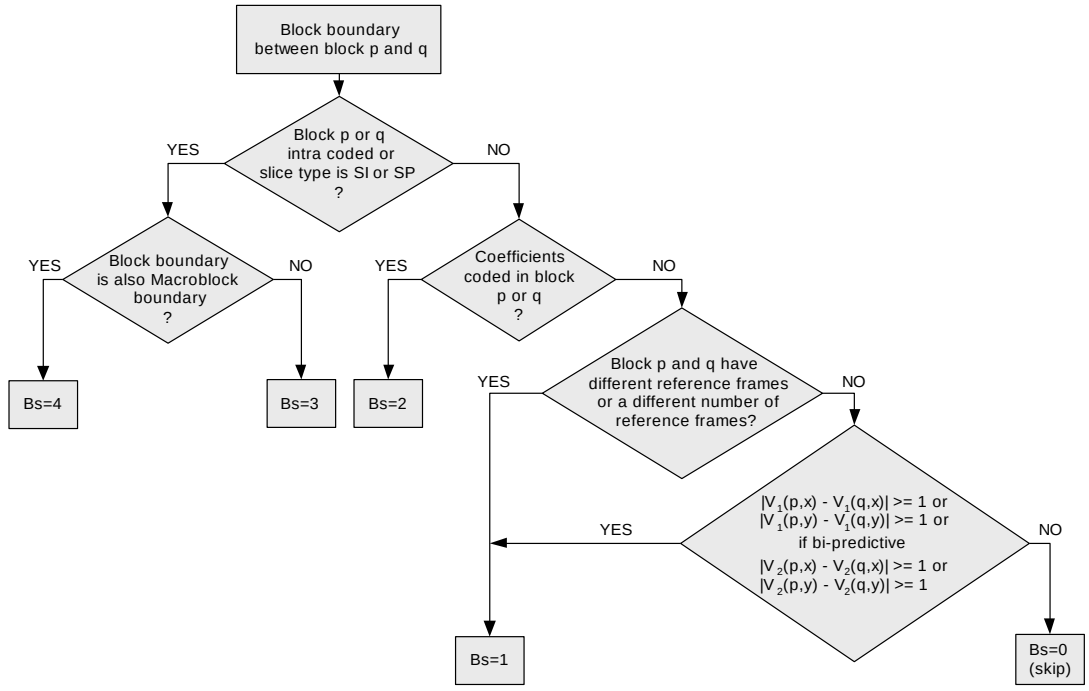


Figure 3.15 – Flow chart for determining the BS ([299])

Figure 3.15 displays the process for computing the BS for the block boundary between two neighbouring blocks p and q, where V1(p,x), V1(p,y) and V2(p, x), V2(p, y) are the horizontal and vertical components of the motion vectors of block p for the first and second reference frames or fields ([299]). The condition equal or superior than 1 signifies that the horizontal distance of the P and Q target blocks in the same reference image is greater than or equal to one integer pixel.

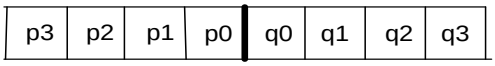


Figure 3.16 – Pixels on either side of a vertical boundary of adjacent blocks P and Q ([299])

The second step is to determine if the filtering operation should be applied ([283],[299]). The filter decision is determined by computing the threshold values α(x) and β(x). Sets of samples across this edge are only filtered if the condition,

$$BS \neq 0 \text{ and } |p0-q0| < \alpha(x) \text{ and } |p1-p0| < \beta(x) \text{ and } |q1-q0| < \beta(x) \tag{3.14}$$

is true. The index x is by default the mean value of the QPs used in P and Q. The average QP value for the two blocks is computed by QPav = (QPp+QPq)>>1. The fundamental idea is that a relatively high absolute variation between pixels near a block boundary implies blockiness and should as a result be decreased ([299]). Nevertheless, if the degree of that difference is so large that it cannot be justified by the coarseness of the quantisation, then the boundary is more likely to display the actual edge in the original picture and should be maintained.

The third step is filtering. Where the filtering conditions are met, filtering occurs. The filter is adapted according to the BS values, pixel positions, and colour components. Further information can be found at ([278],[279],[283],[299],[303]).

3.4.6 *Entropy Coding*

The final step of the video coding process, after all the transform coefficients have been quantised, is entropy coding. The entropy coding algorithm maps the generated syntax elements into binary variable-length binary strings (called codewords), according to a fixed table (called coding table) that can be sent to the NAL Unit in order to be broadcasted. The efficiency of this process depends on whether the length of the binary strings allocated to the syntax elements is matched with their probabilities. The least probable values should be assigned long strings and vice-versa. Although this method has proven to be efficient in terms of computational cost, quite often the compression gain is restricted because the algorithm is not able to adjust the length of the codewords to the varying statistics of the input data. Consequently, researchers have focused on adaptive approaches, in particular the study of adaptive codes that adapt according to the characteristics of the probability distribution of the coded source ([304]).

This is another main difference between H.264/AVC and preceding versions of MPEG. While previous versions use fixed VLC tables that are ‘hard coded’ into the standard, H.264/AVC uses either context-adaptive VLC (CAVLC) tables or context-adaptive binary arithmetic coding (CABAC) for entropy coding ([287],[300],[305]). By integrating context modelling into their

entropy coding framework, both methods offer a high degree of adjustment to the underlying source, though with a different complexity-compression trade-off ([7]).

The CAVLC algorithm is a VLC method with reduced computational complexity. It is applied to the zigzag scanned 4 × 4 block of the transformed coefficients ([299]). The probability that the level of coefficients is zero or +/-1 is very high. CAVLC handles the zero and +/-1 coefficients in a different way to other coefficients. For every block of quantised transform coefficients, the encoder begins by specifying the number of coefficients different to zero and whether there are coefficients equal to ±1 at the end of the scan. Then, based on the number of coefficients, couples (r, l) are coded by writing all the l values first and then the r values afterwards ([279],[286]). It is an effective way of coding the quantised coefficients, adaptively selecting the coding table among a set of possible ones in order to match the signal statistics. This enhances the coding efficiency of the entropy-coding process ([300]).

CABAC is only supported in Main and higher profiles. The CABAC approach is typically more efficient compared to CAVLC but has larger computational complexity. This mode is based on binary arithmetic coding. The efficiency is primarily due to three aspects ([283]). TThe first is that arithmetic coding allows a non-integer number of bits to be allocated per symbol. This is rather advantageous for symbol probabilities higher than 0.5 as variable length codes impose a lower limit of 1 bit/symbol ([305]). This limitation prevents the usage of coding symbols with a reduced alphabet size for coding the residual data. A condensed alphabet would allow a more suitable construction of contexts for exchange among the model probability distributions. A second aspect is that context modelling lets each syntax element have more than one probability model to estimate the conditional probability distributions for distinct local activities. The final aspect is related to the updating process of the probability models. The update process can be performed during the encoding process and thus keeps track of the actual statistics.

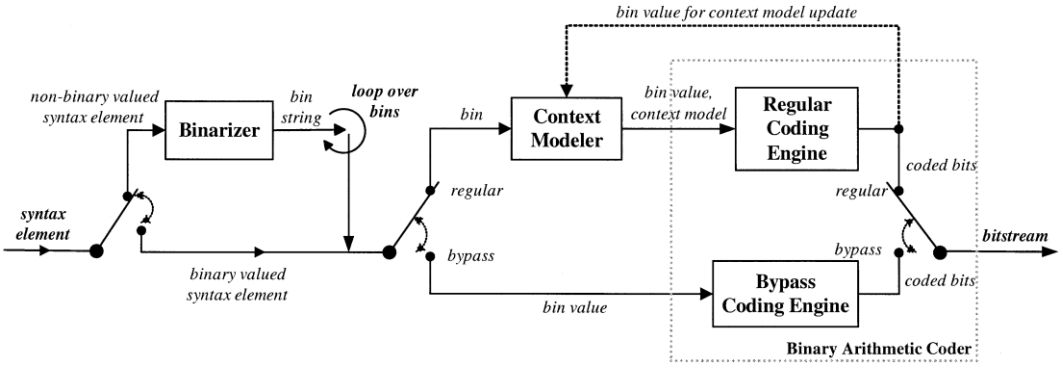


Figure 3.17 – CABAC encoder block diagram ([305])

Figure 3.17 shows the generic block diagram for encoding a single syntax element in CABAC.1 The encoding process consists of three steps: 1) binarisation; 2) context modelling; and 3)

binary arithmetic coding ([305]). Firstly, a non-binary syntax element is processed to produce a unique binary string. This procedure precedes the arithmetic coding step. Next, a probability model is selected for every binary symbol of the binarised string. This model depends on the statistics of recently-coded data symbols. Finally, based on the binary string and the probability model, a binary arithmetic encoder generates the coded bitstream.

3.4.7 *H.264/AVC Profiles and Levels*

The H.264/AVC standard defines several profiles. There are three Profiles in the first version: Baseline, Main, and Extended. Table 3.7 lists the features supported by the different Profiles of H.264/AVC ([279]).

Feature	CBP	BP	XP	MP	HiP	Hi10P	Hi422P	Hi444PP
Chroma formats	4:2:0	4:2:0	4:2:0	4:2:0	4:2:0	4:2:0	4:2:0/4:2:2	4:2:0/4:2:2/4:4:4
Sample depths (bits)	8	8	8	8	8	8 to 10	8 to 10	8 to 14
Flexible macroblock ordering (FMO)	No	Yes	Yes	No	No	No	No	No
Arbitrary slice ordering (ASO)	No	Yes	Yes	No	No	No	No	No
Redundant slices (RS)	No	Yes	Yes	No	No	No	No	No
Data Partitioning	No	No	Yes	No	No	No	No	No
SI and SP slices	No	No	Yes	No	No	No	No	No
B slices	No	No	Yes	Yes	Yes	Yes	Yes	Yes
Interlaced coding (PicAFF, MBAFF)	No	No	Yes	Yes	Yes	Yes	Yes	Yes
CABAC entropy coding	No	No	No	Yes	Yes	Yes	Yes	Yes
8×8 vs. 4×4 transform adaptivity	No	No	No	No	Yes	Yes	Yes	Yes
Quantisation scaling matrices	No	No	No	No	Yes	Yes	Yes	Yes
Separate Cb and Cr QP control	No	No	No	No	Yes	Yes	Yes	Yes
Monochrome (4:0:0)	No	No	No	No	Yes	Yes	Yes	Yes
Separate color plane coding	No	No	No	No	No	No	No	Yes
Predictive lossless coding	No	No	No	No	No	No	No	Yes

Table 3.7 – Features supported in the Profiles of H.264/AVC ([279])

The Baseline profile is aimed at low delay applications, low processing power platforms, and high packet loss environments. The Main profile includes all tools for achieving greater coding efficiency for high bit rate applications (digital storage media and television broadcasting). The Extended profile is geared for error-resilient streaming applications. In addition, there are four High Profiles defined in the Fidelity Range Extension (FRExt) ([279],[295]) which further extend to applications such as content-contribution, content-distribution, and studio editing and post-processing ([285]). The High profile (HP), supporting 8-bit video with 4:2:0 sampling, is

aimed at applications using high resolution such as broadcast and storage applications. The High 10 profile (Hi10P), supporting 4:2:0 video with up to 10 bits of representation accuracy per sample, is also aimed at applications using high resolution, but with a higher bit resolution (higher-quality requirements). The High 4:2:2 profile (H422P), supporting up to 4:2:2 chroma sampling and up to 10 bits per sample, is aimed at professional video applications that use interlace video. Lastly, the High 4:4:4 profile (H444P) that supports up to 4:4:4 chroma sampling at up to 12 bits per sample, additionally includes support of efficient lossless region coding ([286]).

Level No.	Max macroblock processing rate MaxMBPS (MB/s)	Max frame size MaxFS (MBs)	Max decoded picture buffer size MaxDPB (1024 bytes)	Max video bit rate MaxBR (1000 bits/s or 1200 bits/s)	Max CPB size MaxCPB (1000 bits or 1200 bits)	Vertical MV component range MaxVmvR (luma frame samples)	Min compression ratio MinCR	Max number of motion vectors per two consecutive MBs MaxMvsPer2Mb
1	1,485	99	148.5	64	175	[-64,63.75]	2	-
1b	1,485	99	148.5	128	350	[-64,63.75]	2	-
1.1	3,000	396	337.5	192	500	[-128,127.75]	2	-
1.2	6,000	396	891.0	384	1,000	[-128,127.75]	2	-
1.3	11,880	396	891.0	768	2,000	[-128,127.75]	2	-
2	11,880	396	891.0	2,000	2,000	[-128,127.75]	2	-
2.1	19,800	792	1,782.0	4,000	4,000	[-256,255.75]	2	-
2.2	20,250	1,620	3,037.5	4,000	4,000	[-256,255.75]	2	-
3	40,500	1,620	3,037.5	10,000	10,000	[-256,255.75]	2	32
3.1	108,000	3,600	6,750.0	14,000	14,000	[-512,511.75]	4	16
3.2	216,000	5,120	7,680.0	20,000	20,000	[-512,511.75]	4	16
4	245,760	8,192	12,288.0	20,000	25,000	[-512,511.75]	4	16
4.1	245,760	8,192	12,288.0	50,000	62,500	[-512,511.75]	2	16
4.2	522,240	8,704	13,056.0	50,000	62,500	[-512,511.75]	2	16
5	589,824	22,080	41,400.0	135,000	135,000	[-512,511.75]	2	16
5.1	983,040	36,864	69,120.0	240,000	240,000	[-512,511.75]	2	16

Table 3.8 – H.264/AVC Levels and Limitations ([279])

There are several levels for each of the profiles. Each level limits the video resolution, frame rate, HRD bit rate, HRD buffer requirements, and the motion vector range. These limitations are defined in Table 3.8. Entries marked "-" in Table 3.8 denote the absence of a corresponding limit. Levels with non-integer level numbers in Table 3.8 are referred to as “intermediate levels.”

3.5 Summary

This chapter introduces the principal video coding standards. Special attention is devoted to H.264/AVC that adopts many new video coding tools such as intra prediction, integer transform, enhance inter prediction, context-based entropy coding, and deblocking filter. These

new technical developments mean that H.264/AVC achieves a key breakthrough on Rate-Distortion performance. In this work, H.264/AVC is the platform where the proposed techniques will be integrated.

The core of this work focuses on algorithms for controlling the rate control of multiple video sources. Thus, after this overview, the next chapter will be devoted to introducing the non-normative rate control techniques of the major video coding standards.

Chapter 4. HRD Models and Standard Rate Control

Recent video standards such as MPEGx or H.26x standards aim to facilitate interoperability and data exchange among different products or services ([2],[6],[214],[215],[221]). In order to achieve these goals, they specify the requirements imposed on the complete bitstream syntax and decoders.

The standardisation of the decoders enabled independent implementations, from different software and hardware manufacturers, and ensured that those implementations will be interoperable. Flexibility on the decoding side is rather limited while a great deal of flexibility is allowed on the encoding side. This flexibility includes different combinations of picture types in the bitstream, a variable number of reference pictures, buffer size, coding and decoding delay, picture rates and rate control algorithms ([306]). The freedom for determining different parameters on the encoding side allows them to be tuned according to application requirements such as the visual quality for specific channel capacity, random access, or the capacity of a battery in a portable player. These parameters impact on coding efficiency and services such as commercial insertions and multi-channel multiplexing performed in a head-end or a storage server of a digital TV system ([307]). As a result, it is vital to select and recommend procedures for the use of those parameters in order to satisfy the user's expectation regarding the digital television system.

Video standards do not normally define how to perform rate control. Nevertheless, during its development process, algorithms were verified through tests, simulations, and verification models. To allow testing and to perform simulations using a common set of encoder routines, both MPEGx and H.26x set up a sequence of test models as an informative tool (non-normative tool). Each test model normally suggests a rate control method during its development phase, e.g. TM5 for MPEG-2 ([19]), TMN8 for H.263 ([20]), and VM8 for MPEG-4 ([308]), etc. An improved rate control method based on VM8, supporting rate distortion optimisation (RDO), has been adopted by H.264/AVC JM test model ([169],[170]).

Standardising the bitstream syntax and decoding process gives the guarantee that all decoders conforming to the standard will be able to decode any conforming compressed video bitstream according to profile and level capabilities supported by the decoder ([2],[6],[214],

[215],[221],[309]). To accomplish this, it is essential to limit how fast the bitstream data can be sent to a decoder, and the capacity of buffering in the decoder. This is achieved by defining a set of guidelines that takes the form of a flow of the bitstream through a mathematical or hypothetical model of the decoder. This "virtual" decoder is conceptually connected to the output of an encoder and receives the bitstream from the encoder. This model of the decoder is known as the hypothetical reference decoder (HRD) in some standards or recommendations, or the video buffer verifier (VBV) in other standards or recommendations (e.g. MPEG2 Video Buffering Verifier – VBV, H264/AVC- Hypothetical Reference Decoder, etc.) ([307],[310]). Note that contrary to the rate control algorithms, HRD is usually a normative part of video coding standards. These constraints must be enforced by the encoder, and can be assumed by a real decoder or multiplexor to be true. This Chapter reviews HDR models and rate control algorithms associated with the main video standards ([2],[6],[214],[215],[221]).

4.1 HRD Models

A HRD model is composed of a pre-decoder buffer (or VBV Buffer), with size B (in bits), through which compressed data flows with a precisely specified arrival and removal timing schedule, as shown in Figure 4.1 ([311]).

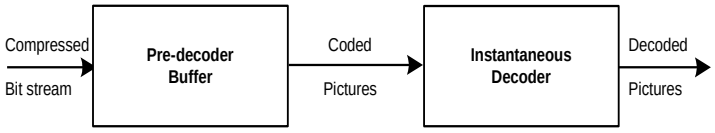


Figure 4.1 – A Hypothetical Reference Decoder

Two additional parameters are important: the channel's peak rate *R* (in bits per second) and the initial start-up delay δ (in seconds). δ can also be represented by the initial decoder buffer fullness *BF* (in bits), since $\delta = BF/R$ ([311],[312],[313]). These parameters characterise resource levels (transmission capacity, buffer capacity, and delay) used to decode a bitstream. The pre-decoder buffer overflows if the buffer becomes full, and more bits are arriving. The buffer underflows if the removal time for a picture occurs before all compressed bits representing the picture have arrived. Initial buffering is fundamental in handling bit rate variations produce by encoding, storing and broadcasting. Otherwise, a streaming system would be susceptible to any type of delay or bit rate variations.

HRDs differ in their means to specify the arrival schedule and removal times, and the rules regarding overflow and underflow of the buffer. The series of removal time and picture size pairs {(tr(n), d(n)), n=0,1,…} is identified as the schedule of a bitstream. The schedule of a

bitstream is inherent to the bitstream, and entirely describes the instantaneous coding rate of the bitstream over its lifetime ([311],[313]).

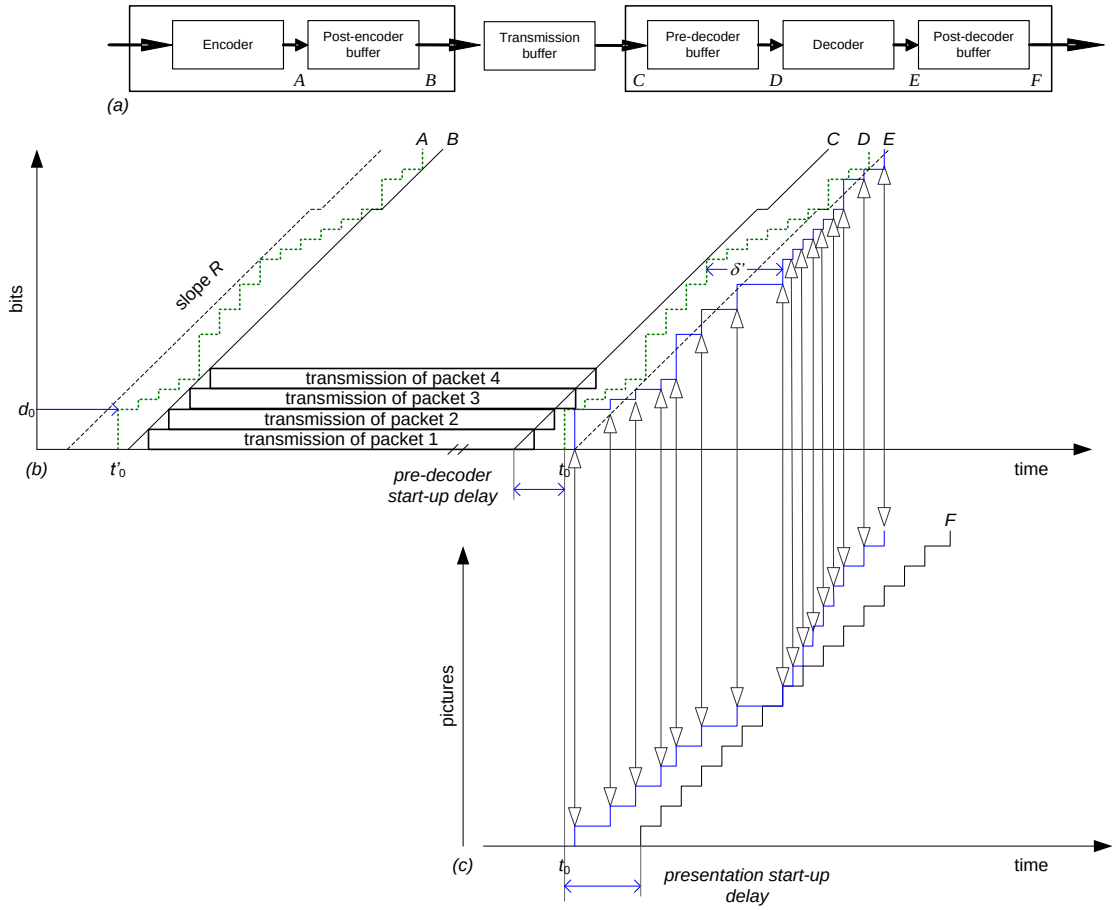


Figure 4.2 – Example of an Encoder-Decoder system buffer

Figure 4.2 (a), exemplifies schedules for the points A, B, C, D, E, and F ([311],[312],[313]). First, a video sequence is encoded and broadcast. It arrives into the pre-decoder buffer at schedule C. The compressed pictures entered the decoder at schedule D. After being decoded the pictures are placed into the post-decoder buffer at schedule E. The pictures are re-ordered and then the uncompressed pictures leave the post-decoder buffer to be displayed (schedule F) ([311],[312],[313],[314]).

In Figure 4.2, the encoder, according to the "encoding" times t'0, t'1, …, instantly place into the post-encoder buffer d0, d1, … bits corresponding to the pictures 0, 1, … (in bitstream order), respectively (schedule A). Bits are removed from the post-encoder buffer, at rate R bits per second (schedule B). The occupancy of the post-encoder buffer (measure in bits) corresponds to the vertical distance between schedules A and B. Bits arrive in the pre-decoder buffer, at schedule C. Then the decoder instantaneously removes the d0, d1, … bits from the pre-decoder buffer, at "decoding" times t0, t1, …, and queues them for decoding (schedule D). The

occupancy of the pre-decoder buffer corresponds to the difference in the buffer between schedule C and schedule D. Schedule D can be shifted in time to avoid the underflow of the buffer, thus minimising the pre-decoder start up delay. The pre-decoder start up delay can still be further minimised by controlling the initial pre-decoder buffer fullness. When the pictures are decoded, they are stored in the post-decoder buffer at schedule E. The maximum computational delay limits the total time that a picture can be stored in the decoder buffer. This parameter also affects the delay between the decoding time of the first picture and its actual presentation time; designated by presentation start up delay (Figure 4.2 (c)).

4.1.1 *Buffering Model in H.263, MPEG-2 and MPEG-4*

The MPEG-2 buffering operation is defined by the VBV (Video Buffer Verifier) and can operate in VBR or CBR modes (Annex C, MPEG-2 [2]). This is a virtual buffer system that allows the encoder to emulate the behaviour of the input buffer in the decoder, during the encoding process. Therefore, the rate control algorithm can have access to relevant information regarding buffer resources. At the decoder, current buffer behaviour follows the completed VBV precisely so that any assumption made at the encoder is exactly adapted at the decoder. During bitstream creation, the VBV fullness must be examined to guarantee that it does not overflow or underflow. There are two key parameters in the VBV model: vbv_buffer_size and vbv_delay. The vbv_buffer_size is the minimum buffer size that has to be allocated at a decoder to decode the corresponding bitstream. The vbv_delay is the time to fill up the VBV buffer to be able to decode without buffer underflow. In MPEG-2 VBV, the removal times are generated considering a fixed frame rate. The exception is film content captured as video (3:2 pull-down). In this circumstance, the removal time of specific pictures is delayed by one field period. The MPEG-2 VBV also defined a low-delay mode. When operating in this mode, no B frames are used and variable frame rate encoding is allowed.

The MPEG-2 VBV supports two operation modes, depending on whether a removal delay (the vbv_delay) is being broadcasted (mode A) or not (mode B) ([2]). In mode A, the rate of arrival in the VBV buffer is calculated for each picture using picture sizes and vbv_delays. An encoder working at CBR can use this mode. Nevertheless, unless the video stream is analysed it is impossible to guarantee that is had been encoded at CBR. An additional drawback is the difficulty of determining the initial rate. In Mode A, the goal is to avoid both buffer underflow and overflow.

In Mode B, no explicit removal delays are broadcasted (vbv_delay is not generated). The arrival rate is constant unless the buffer is full. This approach determines the initial rate. Nevertheless, the arrival schedule may not be causal regarding the real generation of bits. This restricts its use

as support data for a multiplexer. Video encoded data flows into the VBV buffer at the peak rate of the buffer until the buffer is filled, and then it stops. This corresponds to the initial removal time. The following removal times are deferred, with respect to the first, by a fixed frame or field periods. In Mode B, it is impossible for the buffer to overflow, as compressed data inflow stops when the buffer becomes full. Still, it is the encoder's responsibility to prevent underflow. The Mode B model is sometimes referred to as a leaky bucket. This expression derives from the similarity of the encoder to a system that "dumps" water in discrete chunks into a bucket that has a hole in it ([311]). The removal of bits from the encoder buffer relates to the water leaking out of the bucket. A leaky bucket with leak rate R1, bucket size B1, and initial bucket fullness B1–F1 is said to contain a bitstream with schedule {(tr(n), d(n)), n=0,1,…} if the bucket does not overflow under the specific conditions ([311]).

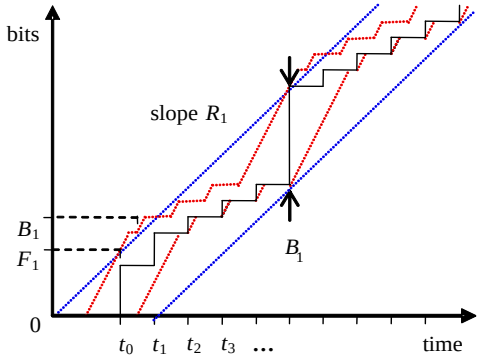


Figure 4.3 – Example of the leaky bucket concept

Consider the example illustrated in Figure 4.3 ([311]). At time t0, d0 bits are placed in the leaky bucket. The bucket starts to pump out video data at the rate R1 bits per second. If the bucket becomes empty, it will stop pumping bits out until more bits are placed into the bucket. At the time ti, i≥ 1, di bits are pumped into the bucket. The bucket keeps on draining bits at a rate of R1 bits per second. The state of the bucket just prior to time ti can be denoted as follows:

$$b_0 = B_1 - F_1 \tag{4.1}$$

$$b_{i+1} = \max\left\{0,\ b_i + d_i - R_1\left(t_{i+1} - t_i\right)\right\} \tag{4.2}$$

The leaky bucket does not overflow if bi + di ≤ B1 for all i ≥ 0. In other words, the leaky bucket holds the bitstream if the graph of the schedule of the bitstream is between two parallel lines with slope R1. The distance between these two lines, measured vertically, is B1 bits (blue line in Figure 4.3). It is possible for the same bitstream to be contained in more than one leaky bucket (red line in Figure 4.3).

The HRD in H.263 was planned for low-delay operation (Annex B, H.263 [215]). At first, the HRD is empty. The encoder and the HDR operate synchronously, at the same clock and picture clock frequency. The HRD buffer is checked at picture clock intervals. If at least one entire coded picture is in the buffer, then all the data for the earliest picture in bitstream order is instantaneously removed (e.g. at t_{n+1} in Figure 4.4).

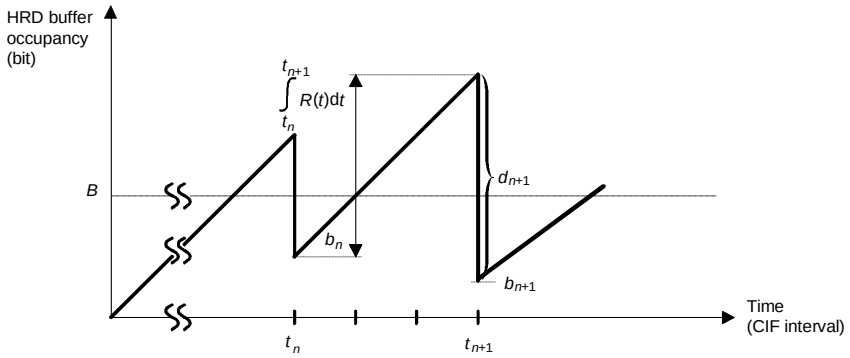


Figure 4.4 – H.263 HRD buffer model ([215])

In H.263, a complete coded picture is one normal I- or P-picture, or a PB-frame or Improved PB-frame (unless the optional Temporal, SNR, and Spatial Scalability mode is in use). Immediately after withdrawing the data, the buffer occupancy must be less than B. The value of B is defined as follows:

$$\mathrm{B} = 4 \times \mathrm{R}_{\max} / PCF \tag{4.3}$$

where PCF is the effective picture clock frequency, and $\mathrm{R}_{\max}$ is the maximum video bit rate during the connection in bits per second ([215]). To meet this requirement the number of bits for the (*n+1*)th coded picture d_{n+1} must satisfy the following condition:

$$d_{n+1} \geq b_n + \int_{t_n}^{t_{n+1}} R(t)dt - B \tag{4.4}$$

where b_n is the buffer occupancy just after time t_n, t_n is the time the *n*th coded picture is removed from the HRD buffer, and *R(t)* is the video bit rate at time *t*. This process differs from removing data associated with a picture at a time explicitly transmitted in the bitstream. It is thus difficult to design systems that display video sequences with precise timing based on this model.

In MPEG-4, VBV (Annex D, MPEG-4 [227]) is defined as a rate buffer model, a complexity model, and a reference memory model. MPEG-4 VBV limits the memory requirements of the coded bit-stream buffering, the processing speed requirements needed for a compliant video

decoder, and the memory requirements of the reconstructed pixel data buffering. If the visual scene is composed of multiple video objects, and if each object contains one or more video objects layers, then the MPEG-4 VBV defines that the video rate buffer model shall be applied independently to each video object layer. An additional requirement is also defined: the total of the sum of the buffer size of each video object layer must not exceed the maximum value defined for the given profile and level ([227]). The question of how to allocate bit rate and buffer size among the various video objects, and, for each video object, among the several video object layers, is outside the scope of the standard.

In MPEG-4 VBV, the buffer is empty in the beginning. Vbv_buffer_occupancy level describes a process to fill the buffer at a data rate to a certain occupancy level before the process of removing data from the buffer starts ([227]). The combination of profile and level define the maximum data rate limit. A picture is instantaneously removed from the VBV buffer at its decoding time and placed in the Video Complexity Verifier (VCV) buffer. The VCV buffer capacity is defined according to the profile and level in use. MPEG-4 VBV and VCV aim to model a real decoder ([227]). VBV denotes buffering prior to decoding, as VCV denotes the decoding process itself.

4.1.2 *HRD Model in H.264/AVC*

Compared with preceding MPEG standards, MPEG-4 establishes some limits on the variability of the number of decoded MB/s, and their complexity ([227]). MPEG-4 video encoders produce bitstreams with a higher variation of the number of macroblocks per second. In MPEG-4, a video scene may be composed of objects whose size varies over time and may be encoded at different video object planes rates. These aspects suggest a higher decoding complexity, compared with previous standards. As a result, MPEG-4 defines a fixed decoding delay (the VCV latency) in order to set some minimum acceptable limits in terms of the decoding capacity of the decoder ([315]).

An MPEG-2 model assumes instantaneous picture decoding at the time of picture decoding. Thus, MPEG-2 decoder implementations can be set to add an arbitrary, non-normative, fixed delay to the decoding process to satisfy the application constraints. The VBV can operate in two modes: CBR and VBR. While MPEG-4 only supports the CBR mode, MPEG-2 supports both modes. The HRD model for H.263 operates in low-delay mode. It resembles the CBR mode of MPEG's VBV, although there are some differences. All these models operate at only one single leaky bucket (R,B,F). H.264/AVC supports multiple leaky buckets in the range 1 to 32 ([310]).

In Annex C of the H.264/AVC standard, the HRD specifies a coded picture buffer (CPB), an instantaneous decoding process, a decoded picture buffer (DPB), and output cropping as shown in Figure 4.5 ([313]). The HRD Coded Picture Buffer represents a means to communicate how the bit rate is controlled in the process of compression ([311],[313]). The arrival and removal time of the coded bits is modelled by CPB. Different from MPEG-2, the H.264/AVC specifies that multiple frames can be used for reference, either from the past or the future in display order. Therefore, the HRD also defines a model of DPB control to guarantee that sufficient memory capacity is used in a decoder for the pictures used as references ([313]).

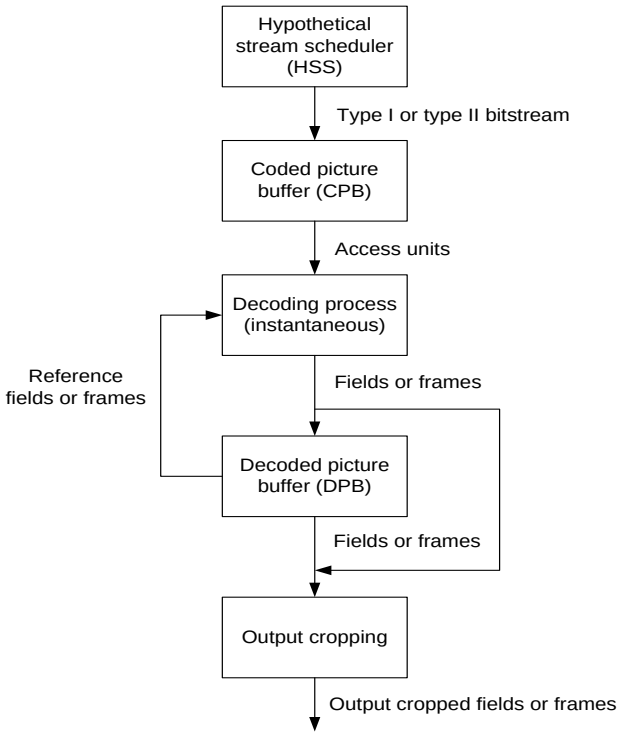


Figure 4.5 – HRD buffer model

The HDR model starts with a video bitstream flow from the HSS into the CPB according to a determined arrival schedule. Second, the video bitstream is removed and decoded instantaneously into the DPB by the decoding process, at CPB removal times. A picture stored in the DPB is removed from the DPB at the DPB output time or when it is marked as "unused for reference." The k^{th} CPB of the HRD is characterized by the pre-decoder peak rate R (in bits per second), the buffer size cpb_size[k] (in bits), the initial CPB removal delay sequence (in seconds), and the picture removal delays for each picture ([313]).

These parameters represent levels of resources (transmission capacity, buffer capacity, and delay) used to decode a bitstream ([311],[313]). In this leaky bucket model, the arrival times of all pictures, after the initial image, are constrained to enter the buffer input at a time interval

greater than the difference regarding the hypothetical encoder processing times between that picture and the first picture.

The operating mode and its main parameters will now be described. The HRD input buffer has a capacity of cpb_size[k] bits (B bits). At the initial stage, the buffer is empty. The lifetime of the coded bits of the picture k in the buffer is represented by the arrival interval, {tai(k), taf(k)} and the removal time tr(k). The arrival interval is defined by the initial arrival time tai(k) and the final arrival time taf(k). The removal time can be computed by (4.5) ([313]):

$$t_r(n) = \begin{cases} \dfrac{initial\_cpb\_removal\_delay}{90000} & n = 0\,(first\,image) \\ t_r(0) + t_c \times \text{cpb\_removal\_delay}(n) & n \neq 0 \end{cases} \tag{4.5}$$

where clock tick tc be specified by

$$\begin{aligned} t_r(0) &= \frac{\text{initial\_cpb\_removal\_delay}}{90000} \\ t_c &= \frac{\text{num\_units\_in\_tick}}{\text{time\_scale}} \end{aligned} \tag{4.6}$$

Due to the arrival time constraint, the initial arrival time of picture n is equal to the final arrival time of the picture n-1, unless that time precedes the earliest arrival time, denoted by the following expression:

$$t_{ai}(n) = \begin{cases} 0 & n = 0 \\ t_{af}(n-1) & t_{af}(n-1) \geq t_{ai,earliest}(n) \\ t_{ai,earliest}(n) & t_{af}(n-1) < t_{ai,earliest}(n) \end{cases} \tag{4.7}$$

where

$$\begin{aligned} t_{ai,earliest}(n) &= t_c \times \text{cpb\_removal\_delay}(n) - t_0 \\ &= t_r(n) - \frac{\text{initial\_cpb\_removal\_delay\_offset}}{90000} \end{aligned} \tag{4.8}$$

Let b(n) be the number of bits associated with picture n. The length of the picture arrival interval can be determined by the time-equivalent of the picture size in bits, at the rate R.

$$t_{af}(n) - t_{ai}(n) \equiv te\left[b(n)\right] = \frac{b(n)}{R} \tag{4.9}$$

In general, the curve of buffer fullness vs. time is given by the following expression ([311],[313]):

$$BF(t) = \sum_n \left[I\left(t_{\mathrm{af}}(n) \le t < t_{\mathrm{r}}(n)\right) \times b(n) + I\left(t_{\mathrm{ai}}(n) < t \le t_{\mathrm{af}}(n)\right) \times be\left(t - t_{\mathrm{ai}}(n)\right) \right] \quad (4.10)$$

Let I(·) denote indicator functions such that I(x) is equal to 1 if x is true and equal to 0 otherwise ([311],[313]).

HRD can also be used to test conformance. ITU-T Recommendation H.264.1 defines conformance specification ([316]). A conformance bitstream needs to be decoded by the reference software decoder specified in ITU-T Rec. H.264.2 ([317]). Furthermore, it needs to satisfy all the requirements for the video layer defined in H.264/AVC, including HRD conformance (based on Annexes C, D and E) ([318]).

Several conformance constraints need to be guaranteed regarding the coded bitstreams ([313]). The first is removal time consistency regarding the precision of the clocks used (90 kHz clock used for initial removal time, tc clock used for subsequent removal time calculations, and bit_rate[k] used for arrival times).

$$\text{initial\_cpb\_removal\_delay} = 90000 \times \left(t_{\mathrm{r}}(n) - t_{\mathrm{r}}(n-1) + t_{\mathrm{ai}}(n)\right) \quad (4.11)$$

A second constraint depends on the coding mode. If the encoder is working at CBR, then data should enter constantly at the CPB such that

$$t_{\mathrm{af}}(n-1) \ge t_{\mathrm{ai,earliest}}(n) \quad (4.12)$$

This time constraint puts a lower boundary on b(n). From (4.12) and (4.8), it follows

$$b(n) = \left(t_{ai,earlist}(n+1) - t_{ai}(n)\right) \times R \quad (4.13)$$

Thus, the lower bound can be computed by

$$B_{underflow}(n) = \max\left(\left(t_{ai,earlist}(n+1) - t_{ai}(n)\right) \times R, 0\right) \quad (4.14)$$

A third constraint and one of the most important is that the buffer must not be allowed to underflow or overflow. Moreover, all pictures, apart from some isolated pictures, must be entirely in the buffer before their removal times. In CBR mode, no gaps in the bit arrival are allowed ([313]). The underflow constraint, $BF(t) \ge 0$ for all t, is guaranteed if the final arrival time of each picture precedes its removal time.

$$t_{\mathrm{af}}(n) \le t_{\mathrm{r}}(n) \quad (4.15)$$

As a result, picture size should be equal to or below the bit-equivalent of the time interval between the start of arrival and the removal time.

$$b(n)=\text{be}\left[te\left(b(n)\right)\right] \le \text{be}\left[t_r(n)-t_{ai}(n)\right] \tag{4.16}$$

where *be(t)* and *te(b)* denote the bit equivalent of a time t and the time equivalent of a number of bits b respectively, and

$$be(t)=t\cdot R;\ te(b)=\frac{b}{R} \tag{4.17}$$

The upper boundary for the picture n is

$$B_{overflow}(n)=\left(t_r(n)-t_{ai}(n)\right)\times R \tag{4.18}$$

As the initial arrival time tai(n) is a function of the sizes and removal delays of previous pictures, the constraint on the b(n) will change over time ([313]). Overflow can be avoided if the buffer fullness curve BF(t) does not exceed buffer size B. The initial pre-decoder removal delay is limited to the maximum value of the time-equivalent of the buffer size, tr(0) ≤ te(B). Furthermore, in normal operation modus, removal delay should not exceed its first values to ensure that no overflow occurs. Regarding the isolated picture size, overflow can be avoided if

$$b(n) \le \text{be}\left[B-t_{ai}(n)\right] \tag{4.19}$$

Rate control must ensure conformance with the Coded Picture Buffer ([313]). In a VBR CPB, the buffer must not overflow or underflow, but gaps may appear in the arrival rate. In order to meet these constraints, the encoder must ensure that for all t, the following inequalities remain true:

$$0 \le BF(t) \le B, \text{ for all } t. \tag{4.20}$$

Using Equation (4.10), this becomes:

$$\begin{aligned} 0 \le BF(t) = \sum_n \Big[& I\left(t_{af}(n) \le t < t_r(n)\right)\times b(n) \\ & + I\left(t_{ai}(n) < t \le t_{af}(n)\right)\times be\left(t-t_{ai}(n)\right)\Big] \le B, \text{ for all } t \end{aligned} \tag{4.21}$$

The buffer fullness BF(t) is a piecewise non-decreasing function of time ([313]), with each non-decreasing interval bounded by two consecutive removal times. Thus, it is enough to guarantee conformance at the interval endpoints ([313]). So, by avoiding underflow at the start of an interval (just after removal of a picture), underflow is fully prevented. The same applies for overflow at the end of the interval, just before picture removal.

4.2 Rate Control Algorithms in Standard Test Models

In recent years, rate control has been one of the main research areas in the field of video coding. Several algorithms and standards have been developed in the field, e.g. MPEG-2, MPEG-4, H.263, and H.264. Rate control usually does not belong to the normative part in video coding standards. Nevertheless, it is the fundamental component for a video encoder to achieve good results, especially when the channel bandwidth and video buffer are limited. Therefore, the video coding standards usually develop rate control algorithms during the standardisation process, like the TM5 (Test Model 5) in MPEG-2 ([19]), TMN8 (Test Model Near-term 8) in H.263 ([20]), VM8 (Verification Model 8) in MPEG-4 Visual ([308]), or JM (Joint Model) in H.264/AVC ([169],[170]). These four rate control schemes will be examined in this section.

4.2.1 *H.263 TMN8 Rate Control Algorithm*

This section presents the TMN8 rate control model, associated to ITU-T H.263, Version 2, the coding standard for low bit rate communication. H.263 Version 2 is designated herein by its working name of H.263+. The test model TMN is a document that describes an effective implementation of an encoder compliant with H.263+. The TMN8 rate control algorithm was designed for low-delay video communications ([20]). The goal was to encode video with good quality for transmitting over a constant bit rate channel while maintain a low buffer delay. In TMN-8, two rate control algorithms are described. Both methods use a buffer regulation scheme in which a target bit rate is chosen, and pictures are skipped until the buffer reaches a limit below the number of bits required to transmit the next picture. Since encoding delays are directly related to buffering fullness, large variations in buffer content will produce undesirable varying delays.

One rate control method works on two levels: frame-layer and macroblock layer. In the frame-layer, a target number of bits per frame is computed. In the macroblock-layer, the quantisation parameter (QP) is adapted to achieve that target. Underlying theory can be found in ([319]). First, the variances of all macroblocks in the motion-compensated picture are computed. Based on these variances, and the remaining bits available for encoding the current picture, model parameters are updated. These parameters are then used to find an "optimal" quantiser for each macroblock. One of the model parameters allows for the weighting of macroblocks based on perceptual importance. The RD model can be described as follows

$$\mathrm{B} = A\left(K \frac{\sigma_k^2}{(2\ \mathrm{QP}_k)^2} + C \right)^2 \tag{4.22}$$

where B is the target number of bits, QP is the quantisation parameter step for the k_th macroblock, A is the number of pixels in a macroblock, K and C are constants, and σ_k^2 is the variance of the luminance and chrominance values in the k_th macroblock.

The test model describes a simple method to calculate this parameter where a macroblock with high spatial activity (higher variance) is assigned a finer quantiser.

An alternate rate control method, described in previous test models, TMN-5, TMN-6, and TMN-7, uses a simpler technique for adapting the quantiser ([320],[321]). In this method, the quantiser is fixed for each macroblock line. The quantiser step size is "manually" adjusted so that the average bit rate for all pictures in the sequence is as close as possible to the one of the target bit rate (8,16 or 32 kb/s) ([20]). When dealing with existent applications, with limited buffer and coding delay, the output bit rate is regulated by adjusting the step size on a macroblock level. The first image is coded as intra, with a fixed quantiser parameter (QP=16). After encoding the picture is finished, the buffer content is adjusted as follows.

$$R / f_{target} + 3x\frac{R}{FR} \quad \text{and} \quad B_{i-1} = \overline{B}. \tag{4.23}$$

with B_{i-1} the number of bits used in the preceding picture, $\overline{B}$ the target number of bits per picture, R the target bit rate, f_{target} the target frame rate, and FR the frame rate of the source material. (typically 25 or 30 Hz). For the remaining pictures, at the beginning of each new macroblock line, a new quantiser parameter is determined as follows:

$$QP_{new} = \overline{QP}_{i-1}\left(1+\frac{\Delta_1 B}{2\overline{B}}+\frac{12\Delta_2 B}{R}\right), \qquad \Delta_1 B = B_{i-1} - \overline{B}, \quad \Delta_2 B = B_{i,mb} - \frac{mb}{MB}\overline{B} \tag{4.24}$$

with $\overline{QP}_{i-1}$ the mean quantiser parameter from the previous picture, mb the current macroblock number, MB the number of macroblocks in a picture, and $B_{i,mb}$ the number of bits spent until the macroblock mb. Only the third term in the formula varies during the picture encoding process. After encoding the entire picture, the buffer content is updated according to:

$$\begin{aligned}
&buffer_content = buffer_content + B_{i,99};\\
&while\left(buffer_content > 3x\frac{R}{FR}\right)\{\\
&\quad buffer_content = buffer_content - \frac{R}{FR};\\
&\quad frame_incr++;\\
&\}
\end{aligned} \tag{4.25}$$

The variable frame_incr indicates the next picture from the source to be encoded. To adjust frame rate, f_{target} and a new $\overline{B}$ are determined at the start of each frame:

$$f_{target} = 10 - \frac{\overline{QP}_{i-1}}{4}; \qquad 4 < f_{target} < 10$$
$$\overline{B} = \frac{R}{f_{target}} \tag{4.26}$$

This process is based on the assumption that encoding is temporarily stopped when the physical transmission buffer is nearly at maximum capacity to avoid buffer overflow. Thus, no minimum frame rate and delay can be guaranteed ([20]). This method is simpler to implement than the one previously described. Nevertheless, it does not provide an accurate quantiser selection making it less effective.

The development of a Lagrangian coder control and parametric choice has led to the creation of a new test model TMN-10 ([322],[323]). TMN-10 has two different methods for determining motion vectors and macroblock coding modes: a low complexity mode using a fast block-matching algorithm, and a high-complexity/high-performance mode using a rate-distortion optimisation algorithm ([322],[323]). In the latter case, first the rate-distortion cost (RDCost) for all possible motion vectors is determined to select the best one having the minimum RDCost (Rate-Constrained Motion Estimation). Thus, the effect of choosing different motion vectors on the overall bit rate and reconstruction distort is assessed. Then, using the best motion vector, the RDcost is determined for all possible modes to select the one that minimizes RDcost (Rate-Constrained Mode Decision). These techniques have been adopted in others test models. They will be briefly described in this section ([324],[325]).

First, for each block or macroblock, a full search on integer-pixel positions is performed to find the "best" motion vector ([322],[323],[324]). The search is further refined by half-pixel refinement. The integer-pixel search is conducted over a window of ±16 pixels, in both the horizontal and vertical directions, relative to the position of the block or macroblock in the current frame. A larger window size is also possible. A Lagrangian formulation is used where distortion is weighted against rate using a Lagrange multiplier. The motion search returns the motion vector that minimizes the following condition

$$J(m, \lambda_{MOTION}) = SAD(s, c(m)) + \lambda_{MOTION} \cdot R(m - p) \tag{4.27}$$

with $m=(m_x, m_y)^T$ the motion vector, $p=(p_x, p_y)^T$ the prediction for the motion vector, and λ_{MOTION} the Lagrange multiplier. The rate term $R(m-p)$ refers the motion information only and is determined by a table-lookup. SAD is given by

$$SAD(s,\ c(m)) = \sum_{x=1,\ y=1}^{B,\ B} \left| s[x,y] - c[x-m_x, y-m_y] \right|, \qquad B = 8, 16. \tag{4.28}$$

where s is the original block or macroblock and c is the predicted block or macroblock. The Lagrangian multiplier λ_{MOTION} is given by

$$\lambda_{MOTION} = 0.92 \cdot QP \tag{4.29}$$

where QP is the macroblock quantisation parameter. The second step is the Rate-Constrained Mode Decision. The constrained optimization problem can be solved by minimizing ([322]).

$$J(s,c,MODE \mid QP, \lambda_{MODE}) = SSD(s,c,MODE \mid QP) + \lambda_{MODE} \cdot R(s,c,MODE \mid QP) \tag{4.30}$$

where QP is the macroblock quantiser, λ_{MODE} is the Lagrange multiplier for mode decision, $MODE$ indicates a mode chosen for a particular macroblock from the set of potential prediction modes available in the standard H.263, and $R(s,c,MODE \mid QP)$ is the number of bits associated with selecting ($MODE$, QP) including the bits for the macroblock header, the motion, and all six DCT blocks. SSD is the sum of the squared differences between the original block s and its reconstruction c given by

$$SSD(s,c,MODE \mid QP) = \sum_{x=1,y=1}^{16,16} \left(s[x,y] - c[x,y,MODE \mid QP] \right)^2 \tag{4.31}$$

and $c\ [x,y,MODE \mid QP]$ and $s[x,y]$ represent the reconstructed and original luminance values. The Lagrangian multiplier λ_{MODE} is given by

$$\lambda_{MODE} = 0.85 \cdot QP^2 \tag{4.32}$$

where QP is the macroblock quantisation parameter.

This second rate control mode aims to solve the problem of optimum bit allocation to the motion vectors and the residual coding. The problem is divided into two parts: motion estimation and code decision. This is, the motion for the INTER is performed first. Then, given the motion vectors, for the rate-constrained mode decision, the overall rate-distortion costs for all the considered macroblocks modes is computed. Regarding rate control, TMN-10 extends

TMN-8 to allocate bits according to picture type: P and B frames. The frame-level in TMN-8 assigns a near-constant target number of bits per P frame. This is a good and useful approach for low-delay video communications. Nevertheless, in scenarios where B frames are used additional techniques are needed, as B frames require fewer bits.

4.2.2 *MPEG-2 Video TM5 Rate Control Algorithm*

The rate control used in the MPEG-2 Simulation Test Model is TM version 5b, known as TM5 ([19]). This model is an improvement of the corresponding algorithm used in the MPEG 1 (Simulation Model 3 - SM 3). The TM5 model is general used throughout research for comparison purposes. The document describes some techniques that were not a matter of standardisation. Some of these techniques were of arguable value but were included to provide a common basis for comparisons ([19]). The Test Model worked as a cookbook for generating bitstreams throughout the collaborative co-experimental phase of MPEG-2 video. The numerous documented tests were an effort to validate the utility of different proposed coding techniques. The last major update of the Test Model document, version 5, took place at the Sydney, Australia meeting of the MPEG working group (WG11) in March 1993.

Video sequences are segmented into GOP units. Within each GOP, a target bit rate is computed, for each frame, using a frame-level bit allocation scheme. On the frame level, the local bit allocation is based on two measurements: deviations from estimated buffer fullness for the macroblock to be encoded and the normalised spatial activity. If the trend of bits generated begins to drift from the estimation, a compensation factor is used to adjust the macroblock quantisation scale (*mquant*). Thus, the quantisation parameters and output buffer content create a closed loop on the macroblock level ([19]). Rate control algorithm determines dynamically the coding parameters. The TM5 rate control algorithm consists of three steps to compute the *mquant* parameter ([19]):

- Bit Allocation: in this step, the number of bits available to code the next picture is estimated. It is performed before coding the picture.
- Rate Control: this step set, using a "virtual buffer," the reference value of the quantisation.
- Adaptive Quantisation: This step modulates the reference value of the quantisation parameter according to the spatial activity in the macroblock to derive the value of the quantisation parameter, mquant. This value is used to quantise the macroblock.

In the first step, global complexity measures are determined according to the picture type: Intra, Predicted and Interpolated. For each frame type, there is a complexity model to estimate the number of bits needed to encode a frame of a given type using a specific quantisation parameter. This model is defined by the following expression

$$B_I \times Q_I = X_I, \quad B_P \times Q_P = X_P, \quad B_B \times Q_B = X_B \tag{4.33}$$

where X_I, X_P, and X_B denote the complexity estimation of a certain type of picture (I, P, or B), whose values are initialized as $X_I = (160 * bit\_rate)/115$, $X_P = (60 * bit\_rate)/115$, and $X_B = (42 * bit\_rate)/115$ respectively. Similarly B_I, B_P, and B_B are the number of bits used for each frame type. Q_I, Q_P, and Q_B are the quantisation parameters used for each frame type. The complexity model is updated after encoding each frame, based on the average quantisation parameter and the number of bits used for that frame. These measures allow different relative weights to be assigned to each picture. The allocation of bits depends also on the buffer fullness and the available bit budget. The bit budget, for a frame, within a GOP, directly depends on the target bit rate, the number of frames of the GOP and depends inversely on the frame rate (Equations (4.34), (4.35) and (4.36)).

$$T_I = \max\left\{\frac{R}{1+\frac{N_p X_p}{K_p X_i}+\frac{N_b X_b}{K_b X_i}}, \frac{bit\_rate}{8\times picture\_rate}\right\} \tag{4.34}$$

$$T_P = \max\left\{\frac{R}{N_p+\frac{N_b K_p X_b}{K_b X_p}}, \frac{bit\_rate}{8\times picture\_rate}\right\} \tag{4.35}$$

$$T_B = \max\left\{\frac{R}{N_b+\frac{N_p K_b Xp}{K_p X_b}}, \frac{bit\_rate}{8\times picture\_rate}\right\} \tag{4.36}$$

where T_I, T_P, and T_B are the target number of bits for the corresponding frame type, N_p and N_b are the numbers of P pictures and B pictures remaining in the current GOP in the encoding order respectively, K_p and K_b are constants with default values 1.0 and 1.4 respectively, and R

is the remaining number of bits assigned to the GOP. After encoding a picture, R is updated by subtracting the number of bits used from the bit budget of the GOP, as follows

$$R = R - B_{i,p,b}, \tag{4.37}$$

where $B_{i,p,b}$ is the number of bits generated in the picture just encoded (picture type may be I, P or B). Before start encoding the first picture in a GOP, considering the exceeding bits R_{prev} of previous GOP, initial value of R is given by

$$R = G + R_{prev} \tag{4.38}$$

where G = bit_rate*N/picture_rate, and N is the number of pictures in the GOP. For the first GOP Rprev= 0.

The second step is the rate control. Within a picture, it is estimated the number of bits needed to encode a macroblock and the quantisation step size adjusted. If a significant difference occurs between the target bits (computed before beginning encoding the picture) and the generated bits when data is coded, then rate control adjusts the bit allocation process. If the virtual buffer begins to overflow, the macroblock quantisation step size is increased. This will result in a decrease of the number of coded bits in subsequent coded macroblocks. Likewise, if the level of the virtual buffer approaches the underflow level, a reverse process takes place, and the quantisation step size is decreased. The buffer fullness, d_m^n, is estimated by

$$d_m^n = d_0^n + B_{m-1} - \frac{(m-1)}{MB\_CNT} T_n, n = i, p, b, \tag{4.39}$$

d_m^i, d_m^p, and d_m^b are the virtual buffer's occupancy at macroblock m for each picture type. The final fullness of the virtual buffer when a picture is completly encoded (d_m^i, d_m^p, d_m^b : m=MB_CNT) is used as the initial value for the next picture. B_{m-1} is the number of bits generated by encoding all macroblocks in the picture up to an including macroblock m-1. MB_CNT is the number of macroblocks in a frame. Next, the quantisation step size for the macroblock m, Q_m, is computed as follows

$$Q_m = \left(\frac{d_m \times 31}{r} \right) \tag{4.40}$$

where the "reaction parameter" r is given by

$$r = 2 \times \frac{bit\_rate}{picture\_rate} \tag{4.41}$$

According to the Equation (4.40), the quantisation parameter is proportional to virtual buffer fullness. Note that this virtual buffer has no relation with the HRD buffer. It is a virtual buffer,

at the encoder side, used to compute values of the quantisation parameter linearly ([282]). This procedure is illustrated in Figure 4.6 ([19]).

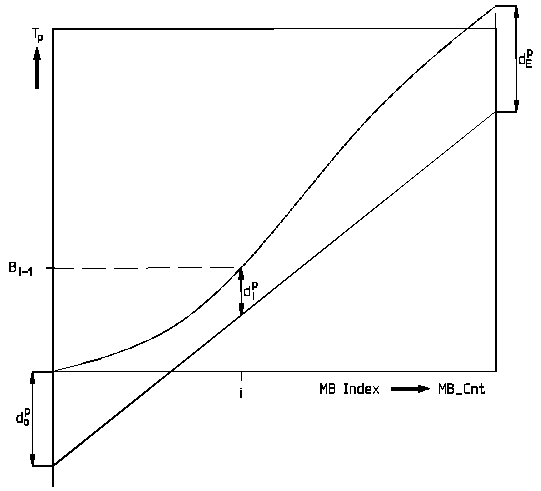


Figure 4.6 – Rate Control for P-pictures

The fullness of the virtual buffer for the last MB is used for encoding the next picture of the same type as the initial fullness. In experiments of the TM5 development process, the maximum size of the virtual buffer was taken as "two times" the average picture bit budget ([282]). The initial value for the virtual buffer fullness is given by:

$$d_0^i = 10 \times \frac{r}{31},\ d_0^p = K_p \times d_0^i,\ d_0^b = K_b \times d_0^i \tag{4.42}$$

In the final step, the macroblock quantisation step size is modulated by a local activity measure. For example, consider a macroblock with very complex and strong spatially-varying intensities. In this case, a given stimulus will be harder to notice (spatial masking effect). On the other hand, in very smooth areas, equivalent changes may be quite visible. The local activity of a macroblock corresponds to the minimum value of the luminance variances of the four 8x8 blocks that compose the macroblock.

$$act_j = 1 + \min\left(vblk_1, vblk_2, ..., vblk_8\right) \tag{4.43}$$

where

$$\begin{aligned} vblk_n &= \frac{1}{64} \times \sum_{k=1}^{64} \left(P_k^n - P_{\_mean_n}\right)^2 \\ P_{\_mean_n} &= \frac{1}{64} \times \sum_{k=1}^{64} P_k^n \end{aligned} \tag{4.44}$$

and P_k^n are the samples values in the nth original 8x8 block. The value of the macroblock spatial activity is normalised as follows:

$$N\_act_j = \frac{\left(2 \times act_j\right) + avg\_act}{act_j + \left(2 \times avg\_act\right)} \tag{4.45}$$

where avg_act is the average value of act_j in the last encoded frame. For the first frame, avg_act is set to 400. TM5 finds the value of the quantisation parameter $mquant_j$ of MB_j as follows:

$$mquant_j = Q_j \times N\_act_j \tag{4.46}$$

The value of $mquant_j$ is clipped to the range [1,31]. The TM5 is based on the following assumptions:

1. The distortion D increases linearly with the quantisation parameter q;
2. To maintain constant video quality, *mquant* values for I, P, and B frames (*mquantI*, *mquantP*, *mquantB)*, are related by constant *kP* and *kB*;
3. The coding bit rate R is inversely proportional to the distortion D. In other words, R x D = constant.

Human vision is not so sensitive to the quantisation noise in pictures with a high amount of movement or spatial detail, while the degree of sensibility increases in uniform areas. TM5 aims to allocate more bits in the smooth areas and fewer bits for the active areas. Note that the third assumption is based on an extremely simplified RD model. Thus, the performance of the TM5 rate control algorithm suffers from accuracy problems.

4.2.3 *MPEG-4 VM8 Rate Control Algorithm*

In 1996, the MPEG meeting that took place at Munich defined the first MPEG-4 Video Verification Model (VM). The MPEG-4 VM presented a novel approach on content based data representation ([275]). In this approach, a scene is viewed as a composition of Video Objects (VO), an arbitrarily shaped time-variable visual entity that can be individually manipulated and combined with other similar entities to produce a scene ([326],[327]). Each object in the scene is coded independently, and a coded scene corresponds to the multiplexing of the different video bitstreams resultant of the video objects (VOs) that composed the scene. When encoding an object, the amount of bits necessary to encode the shape information also needs to be considered. Additionally, each VO may be encoding at a different frame rate. VM8 provides

support scalable rate control (SRC) for diverse bit rates, spatial resolutions and temporal resolutions and various coders (e.g. DCT and wavelet) ([308],[328],[329]).

The image sequences in MPEG-4 that are to be encoded in each VOP layer are, in general, considered to be entries of arbitrary shape. A single VOP layer is supported as a special case. At the core of object-based video, coding is the question on how to allocate efficiently bits among different objects in a scene. VM has evolved through time by core experiments ([327]). This algorithm became stable in the MPEG-4 Video VM 8.0 (VM8) ([308],[315]). VM provides informative description of a rate control scheme composed of three algorithms: Frame Rate Control, Multiple Video Object Rate Control, and Macroblock Rate Control. Thus, a rate control algorithm can be applied to single video object (SVO) and independent multiple video object (MVO) encoding ([315]), and rate control is implemented in both the picture and macroblock level where a second-order rate-distortion model is used for target bit allocation ([327],[330]). The rate control model is formulated as ([308]):

$$R(Q)=\left(X_1\times\frac{1}{Q}+X_2\times\frac{1}{Q^2}\right)\times Ec=\frac{a}{Q}+\frac{b}{Q^2} \tag{4.47}$$

where R is the bit rate, Ec the encoding complexity expressed by the mean absolute difference (MAD) between the VOP being encoded and its reference, Q the quantisation parameter, and X_1 and X_2, or a and b, the model parameters. Using coding statistics such as the average quantisation parameter and the total coding bit rate from previous frames, for each type of picture, model parameters can be estimated for the current frame by a linear regression. Model parameters a and b can be determined as follows:

$$b=\frac{n\sum_{i=1}^{n}R_i-\left(\sum_{i=1}^{n}Q_i^{-1}\right)\left(\sum_{i=1}^{n}Q_iR_i\right)}{n\sum_{i=1}^{n}Q_i^{-2}-\left(\sum_{i=1}^{n}Q_i^{-1}\right)^2} \tag{4.48}$$

$$a=\frac{\sum_{i=1}^{n}Q_iR_i-b_kQ_i^{-1}}{n} \tag{4.49}$$

where n is the number of frames encoded, Q_i and R_i the values of the current average quantisation parameter and bit count in the past ([282]).

There are four steps in the MPEG-4 VM8 frame rate control algorithm ([315],[327],[331]). The first step is initialization. It consists of determining the model parameters, the number of bits

remaining for encoding the sequence (or segment), R_r, and the average number of bits to remove from the encoder buffer per encoding time instant R_p. These parameters are computed as follows

$$X_1 = \frac{R_S \times N_S}{2}, X_2 = 0 \tag{4.50}$$

$$R_r = T_S \times R_S - R_f, \quad R_p = \frac{R_r}{N_r} \tag{4.51}$$

with T_S the number of seconds remaining to encode the sequence (or segment), R_S the target bit rate for the sequence (or segment), R_f the number of bits used to encode for the first frame, N_S the distance between encoded frames, and N_r the number of P frames remaining for encoding.

In the second step, target bit allocation is performed before encoding. The target number of bits to be used during the encoding of the current frame, T, depends on the bits available and the last encoded frame bits, and is determined by

$$T = \max\left[\frac{R_S}{30}, (1-\alpha) \times \frac{R_r}{N_r} + \alpha \times S\right] \tag{4.52}$$

with S the number of bits used for encoding the previous frame, and $\alpha = 0.05$. If the last frame was complex, then it used a high number of bits. Thus, the current frame should be assigned more bits but fewer bits are available for encoding. A weighted average (α)reflects the trade-off between these two factors. To guarantee minimal quality, a lower target bit rate is used $R_S/30$. The target bit rate is adjusted according to the buffer status to avoid both overflow and underflow.

$$T = T \times \frac{\left(\mathrm{B} + 2 \times \left(\mathrm{B}_S - \mathrm{B}\right)\right)}{\left(2 \times \mathrm{B} + \left(\mathrm{B}_S - \mathrm{B}\right)\right)} \tag{4.53}$$

with B the current video buffer level, and B_S the buffer size. If, even after this adjustment, the number allocated can still lead to a potential violation of the VBV constraints, a more extreme adjustment is performed.

$$T = \begin{cases} \max\left[\frac{R_S}{30}, \beta B_S - B\right] & \Leftarrow B + T > \beta B_S \quad \textit{to avoid overflow} \\ R_P + \left(1-\beta\right) B_S - B & \Leftarrow B + T - R_P < \left(1-\beta\right) B \quad \textit{to avoid underflow} \end{cases} \tag{4.54}$$

where $\beta = 0.9$.

In the third step, the value of the quantisation parameter Q_c, before encoding the current frame is estimated. First, T is further tuned to guarantee that more bits are used to encode the residual texture data than the number of header and motion vector bits used in the previous frame (H_p).

$$\mathrm{T} = \mathrm{Max}\left[\frac{R_p}{3} + H_p, T\right] \tag{4.55}$$

Using Equation (4.55) in Equation (4.47), Q_c can be solved based on the model parameters as follows (Equation (4.56))

$$\begin{aligned} &if \quad \left(X_2 = 0\right) \vee \left(\left(X_1 E_c\right)^2 + 4X_2 E_c\left(T - H_p\right)\right) < 0 \\ &\qquad Q_c = X_{11}\frac{E_c}{\left(T - H_p\right)} \\ &else \quad Q_c = \frac{2X_2 E_c}{\sqrt{\left(X_1 E_c\right)^2 + 4X_2 E_c\left(T - H_p\right)} - X_1 E_c} \end{aligned} \tag{4.56}$$

with X_{11} the model parameter for the fall back first-order model ([315],[327]). The quantisation parameter must remain in the range of 1 and 31. Furthermore, Q_c can only vary within 25% of the previous Q_c to maintain a VBR quality ([327]). Thus, the value of the quantisation parameters is clipped to satisfy these conditions. After this step, the image is encoded.

The fourth and final step, is designated as post encoding. In this step, model parameters are updated based on the statistics generated during the encoding of the current frame. The level of buffer fullness is updated by adding to it, the number of bits used during the encoding process, and by subtracting R_p. The value of the remaining bits is updated by deducting from it, the number of bits that were spent. The rate distortion model is updated based on the encoding results of the current frame. A first estimate of the model parameters is computed using a sliding window to select the number of data points used to estimate the model parameters ([308]). The value of w is computed as follows $w[i] = \min\left[total\_data\_number, w_{\max}\right]$

$$\begin{cases} w[k] = ceil\left(\dfrac{MAD_k}{MAD_{k-1}} \times [k-1]\right) & : if\ \left(MAD_{k-1} > MAD_k\right) \\ w[k] = ceil\left(\dfrac{MAD_{k-1}}{MAD_k} \times w[k-1]\right) & : otherwise \end{cases} \tag{4.57}$$

with k the time instant basic unit, $W_{\max}$ the maximum value of sliding-window (default value is 20). If the encoding frame complexity varies significantly, the window size can be reduced with more recent data points. One example is after a scene change occurs. In order to improve the estimation of the model parameters, the outlier data points are rejected by using a statistic's measurement as follows.

$$std = \sqrt{\frac{1}{w}\sum_{k=0}^{w}\left(\frac{X_1 \times MAD_k}{Q_k} + \frac{X_2 \times MAD_k}{Q_k^2} - R_k\right)^2} \tag{4.58}$$

$$e(k) = \left|\frac{X_1 \times MAD_k}{Q_k} + \frac{X_2 \times MAD_k}{Q_k^2} - R_k\right| \tag{4.59}$$

with w the number of past basic unit, and R_k the actual coded bits of the k-th basic unit. If $\mathrm{e}(k) > \mathrm{std}$, basic unit k is rejected, and thus not included in model parameters estimation of X_1 and X_2. A second estimate is then performed using a linear regression technique.

Extensions to multiple VO (MVO) and other improvements were addressed in the Core Experiment Q2 and were presented in VM8 ([332]). For I-frame and first P-frame coding, macroblock-based quantisation parameters are determined according to an algorithm based on human visual sensitivity (HVS) ([333]). For the remaining P-frames, solutions have been proposed in the literature ([334],[335]). In VM8 MVO rate control, each object maintains its own set of parameters ([308]). First, an initial target estimate is made for each object. Based on the buffer fullness, the total target is adjusted and then distributed in proportion to the size of the object, the motion that the object is experiencing, and the value of the metric MAD (mean absolute difference). New quantisation parameters, with the individual targets and the quadratic rate-distortion model, are computed for each video object. In addition, an algorithm is described to control shape rate.

The target bit allocation and rate control algorithm in MPEG-4 VM8.0 is similar to the rate control mechanism proposed in the TM 5. The main difference is that TM5 uses a linear rate-distortion model, and VM8 uses a quadratic model. The VM8 rate control algorithm often suffers from severe performance degradation at scene changes, because VM8 is based on the assumption that adjacent pictures of equal type have the same rate and distortion characteristics. Likewise, the VM8 algorithm also suffers from relatively large control error due to the limited accuracy and robustness of its rate model (Eq. (4.47)) ([21],[319]). A more detailed description, containing advances and trends in terms of rate control for object-based video coding can be found ([336],[337],[338]).

4.2.4 *H.264/AVC JM Video Rate Control*

Existing studies indicate that H.264/AVC brings major improvement in coding performance compared with preceding coding standards ([298]). H.264/AVC presents many new features that represent challenges to operative encoder control. Like previous standards, rate control is not a part of the H.264 standard. Similar to preceding standards, non-normative guidance has been issued, and is known as the Joint Model (JM) ([169],[170],[339],[340]). The rate control in JM incorporates a Rate Distortion Optimized motion estimation and mode decision (also referred to as RDO). This is a major contribution to the high coding efficiency of H.264/AVC compared with previous video compression standards. Nevertheless, this innovation increases the complexity of the rate control process due to the inter-dependency between the RDO and rate control. The rate control algorithm can access the exact coding characteristics only after completing the intra/inter prediction process. With this information, it determines the quantisation parameter. Thus, rate control cannot access the exact coding characteristic in advance. This problem of deciding which parameter should be first determined is sometimes referred in the literature as the "chicken and egg" dilemma ([341],[342],[343]). To avoid this dilemma, a two-pass scheme was proposed in JVT-D030 ([344]). In each pass, a TM-5-alike method is used. Because of the simplified R-D function that was selected, this approach fails to achieve an accurate and robust rate control ([339]). In addition, the level of complexity is increased due to the two-pass strategy. To overcome these drawbacks, JVT-G012 ([339]) was proposed and selected to be include in the JM. In JVT-G012, a linear MAD model predicts the coding complexity, and a MPEG-4 Q2 function employed to estimate the quantisation parameter.

The rate control in JM is based on the hypothesis that a GOP is formed. We should be aware that a GOP is not a mandatory concept in H.264. In fact, there is no formal GOP structure since referencing order is decoupled from coding order in H.264 ([282]). A coded video sequence in H.264/AVC consists of a series of access units. It starts with an Instantaneous Decoding Refresh (IDR) access unit composed of one or more IDR slices, a special type of Intra coded slice. The presence of an IDR access unit indicates that all the following coded pictures, in decoding order, can be decoded without inter prediction from any picture decoded preceding the IDR picture. The next video frames or fields are coded as slices. The size of the slices varies according to the encoding applications. The most common slice size is one slice per coded picture. A second scenario is to divide a picture in N slices, with each slice having roughly the same number of bits. In this case, the number of macroblocks will vary. This could be useful if, for example, each slice is mapped to a fixed-size network packet. A third scenario is to divide a picture in N

slices, each containing the same number of macroblocks. In this case, the number of bytes will vary depending on the picture area characteristics (amount of motion and detail).

primary_pic_type	slice_type values that may be present in the primary coded picture
0	I
1	I, P
2	I, P, B
3	SI
4	SI, SP
5	I, SI
6	I, SI, P, SP
7	I, SI, P, SP, B

Table 4.1 – Meaning of primary_pic_type ([6])

Slice_type	Contains macroblocks types	Notes
I (I slice)	I only	Intra prediction only.
P (P slice)	I and/or P	Intra prediction (I) and/or prediction from one reference per macroblock partition (P).
B (B slice)	I, P and/or B	Intra prediction (I), prediction from one reference (P) or biprediction, i.e. prediction from two references (B)
SP (SP slice)	P and/or I	Switching P slice
SI (SI slice)	SI	Switching I slice

Table 4.2 – Name association to slice_type ([6])

Compared to previous standards, H.264/AVC does not specify the picture's type as picture type I/P/B since each picture is composed of slices with different slice types (Table 4.1, Table 4.2) ([6],[282]). Nevertheless, in the literature, these names have been used ([307]). Typically, an I-picture has been associated with a picture containing only intra coded slices and whose macroblocks do not use inter prediction. A “special” type of Intra pictures is defined in H.264/AVC: the IDR picture. The P-picture, also referred to as predicted pictures, have been associated to pictures composed of macroblocks coded as Intra or coded as Inter but with only one motion vector. This type of “P-picture” differs from MPEG-2 as the reference frames can be from either the past or the future relative to the current frame. Pictures designated as “B-pictures” refer to pictures that support both Intra and Inter coding mode. In the case of Inter code mode, macroblocks can be encoded with up to two motion vectors. They also differ from classic MPEG-2 B picture type. H.264/AVC supports the prediction where both references can be in the future or in the past. In addition, parts of the image coded as B macroblocks can be used as a reference for temporal prediction by other pictures in the past or in the future. In the standard, the characteristics of each coded picture are declared with primary_pic_type whose value specifies the kind of slices are in the coded picture (Table 4.1) ([6]). For example, a coded picture can be constituted by I, P and B Slices, and a coded B slice can be composed of I, P and

B macroblocks. As mentioned before, JM uses a quadratic R-D model with a linear regression method similar to MPEG-4. As pictures are composed of slices with varied slice types, a signal deviation measure, σ_{signal}, is introduced to represent the distortion measure instead separately tuned model parameters for different picture type. Thus, the number of target bits can be adjusted according to picture characteristics using the signal deviation σ_{signal}. Pictures with lower values of signal deviation should receive fewer bits, and pictures with higher values of signal deviation should be allocated more bits. At slice level, a similar approach is followed to obtain uniform rate control. In the JM model, σ_{signal} is determined using MAD, similar to the Encoding Complexity, in MPEG-4, also expressed in MAD.

The video sequence ends when the transmission is complete, or when a new IDR slice is generated. The JM model consists of three components: a GOP level rate control, a picture level rate control and the optional basic unit level rate control (when the basic unit is not defined as a frame). A basic unit is defined as a group of successive macroblocks in the same frame. For example, a basic unit can be a macroblock, a slice, a field, or a frame. GOP level rate control computes the initial quantisation parameter of the instantaneous decoding refresh (IDR) of the picture and that of the first stored picture, and the total bits for the remaining frames in the GOP. In H.264/AVC, the IDR picture is the first coded picture of the video sequence to be coded ([339],[340]). It contains only I or SI slices. Notice that the GOP level rate control allows the restraining of the variation of the quantisation parameters to a smaller scale between neighboring GOPs. Thus, the difference in quality is not very noticeable. The picture layer rate control consists of two stages: pre-encoding and post-encoding stages. The goal of the pre-encoding stage is to calculate the quantisation parameter for each picture. In the post-encoding stage, model parameters and the coefficients in the quadratic R-Q model, are updated. The basic unit-layer rate control is similar to the rate-control of the frame-layer ([169],[170],[339],[340]). This algorithm is now presented in more detail.

4.2.4.1 GOP level rate control

In this phase, it is estimated, for each GOP, the total number of remaining bits for the remaining frames of the GOP and it is initializes the QP of the instantaneous decoding refresh (IDR) frame ([339]). After encoding the j^{th} picture in the i^{th} GOP, the sum of the bits for the remaining pictures in the i^{th} GOP can be computed as follows

$$B_i(j) = \begin{cases} \dfrac{R_i(j)}{f} \times N_i - V_i(j) & j = 1 \\ B_i(j-1) + \dfrac{R_i(j) - R_i(j-1)}{f} \times (N_i - j + 1) - b_i(j-1) & j = 2, 3, ..., N_i \end{cases} \tag{4.60}$$

where f is the predefined coding frame rate. N_i is the total number of pictures in the i[th] GOP. $R_i(j)$ and $V_i(j)$ are the instant available bit rate and the occupancy of the virtual buffer, respectively, when the j[th] picture in the i[th] GOP is coded, and $b_i(j\text{-}1)$ is the actual generated bits in the (j-1) [th] picture ([339]). In the case of the dynamic channels, $R_i(j)$ may vary at different frames and GOPs. Nevertheless, in the case of constant bit rate, $R_i(j)$ is always equal to $R_i(j\text{-}1)$ and the equation can be simplified as follows

$$B_i(j) = B_i(j-1) - b_i(j-1) \tag{4.61}$$

After encoding a picture, the virtual buffer level $V_i(j)$ is updated as

$$\begin{aligned} V_i(1) &= \begin{cases} 0 & i = 1 \\ V_{i-1}(N_{i-1}) & other \end{cases} \\ V_i(j) &= V_i(j-1) + b_i(j-1) - \frac{R_i(j-1)}{f} \qquad j = 2,3,...,N_i \end{aligned} \tag{4.62}$$

For the first GOP, an initial quantisation parameter $QP_1(1)$ is set for the IDR picture using Equation (4.63)

$$QP_1(1) = \begin{cases} 40, & bpp \le l1 \\ 30, & l1 < bpp \le l2 \\ 20, & l2 < bpp \le l3 \\ 10, & bpp > l3 \end{cases} \tag{4.63}$$

where

$$bpp = \frac{R_1(1)}{f \times N_{pixel}} \tag{4.64}$$

and N_{pixel} is the number of pixels in a picture. $l1 = 0.15, l2 = 0.45, l3 = 0.9$ are the values recommended for QCIF/CIF images, and $l1 = 0.6, l2 = 1.4, l3 = 2.4$ are the values recommended for picture size greater than CIF ([169],[170],[339]). For the other GOPs, the quantisation parameter is updated with Equation (4.65):

$$QP_i(1) = \max\left\{QP_{i-1}(1) - 2, \min\left\{QP_{i-1}(1) + 2, \frac{SumPQP(i-1)}{N_p(i-1)} - \min\left\{2, \frac{N_{i-1}}{15}\right\}\right\}\right\} \tag{4.65}$$

where $N_P(i\text{-}1)$ is the total number of stored pictures in the (i-1)[th] GOP, and $SumPQP(i\text{-}1)$ is the sum of the average value of the picture quantisation parameters for all stored pictures in the (i-1)[th] GOP. This value can be further adjusted by:

$$QP_i(1) = QP_i(1) - 1 \quad if \quad QP_i(1) > QP_{i-1}(N_{i-1} - L) - 2 \tag{4.66}$$

where $QP_{i-1}(N_{i-1} - L)$ is the quantisation parameter of the last stored picture in the previous GOP, and L is the number of successive unstored pictures between two stored pictures ([169],[170]).

4.2.4.2 Picture level rate control

In the picture level rate control, two stages are involved: pre-encoding and post-encoding. In the pre-encoding stage, the quantisation parameter of each picture is calculated applying different methods for stored and unstored pictures. In the second case, a quantisation parameter is computed with a linear interpolation of the quantisation parameters between two adjacent stored pictures. Consider the case of two pictures, j^{th} and $(j+L+1)^{th}$, in the i^{th} GOP. When there is only one unstored picture between these two stored pictures (L=1), the quantisation parameter for the $(j+1)^{th}$ unstored picture, $QP_i(j+1)$ is computed by the following expression:

$$QP_i(j+1) = \begin{cases} \dfrac{QP_i(j) + QP_i(j+2) + 2}{2} & \text{if } QP_i(j) \neq QP_i(j+2) \\ QP_i(j) + 2 & \text{Otherwise} \end{cases} \tag{4.67}$$

When L>1, the quantisation parameter for the $(j+k)^{th}$ $(1 \le k \le L)$ unstored picture is computed by the following expression

$$\begin{aligned} QP_i(j+k) &= QP_i(j) + \alpha + \\ &+ \max\left\{ \min\left\{ \frac{(QP_i(j+L+1) - QP_i(j)}{L-1},\ 2(k-1) \right\}, -2(k-1) \right\}, k = 1,..,L \end{aligned} \tag{4.68}$$

In the above equation, α is a function of the difference between the quantisation parameter of the first unstored picture and $QP_i(j)$, and is given by

$$\alpha = \begin{cases} -3 & QP_i(j+L+1) - QP_i(j) \le -2 \times L - 3 \\ -2 & QP_i(j+L+1) - QP_i(j) = -2 \times L - 2 \\ -1 & QP_i(j+L+1) - QP_i(j) = -2 \times L - 1 \\ 0 & QP_i(j+L+1) - QP_i(j) = -2 \times L \\ 1 & \mathrm{QP_i}(j+L+1) - QP_i(j) = -2 \times L + 1 \\ 2 & \text{Otherwise} \end{cases} \tag{4.69}$$

In the final step, the quantisation parameter $QP_i(j+k)$ is bound to a value between 0 and 51. In the case of the stored pictures, the value of the target bits for each stored picture in the current GOP is computed. This value depends on several factors such as the target buffer level, the

frame rate, the available bandwidth, and the actual buffer occupancy. A target buffer level is predefined for each stored picture according to the average picture complexity, and the value of the coded bits of the first IDR picture and the first stored picture. After having coded the first stored picture in the i[th] GOP, the initial value of target buffer level is set to $S_i(2) = V_i(2)$ (the occupancy of the virtual buffer). The target buffer level for the subsequent stored picture is obtained by

$$S_i(j+1) = S_i(j) - \frac{S_i(2)}{N_p(i)-1} + \frac{\bar{W}_{p,i}(j)\times(L+1)\times R_i(j)}{f\times(\bar{W}_{p,i}(j)+\bar{W}_{b,i}(j)\times L)} - \frac{R_i(j)}{f} \tag{4.70}$$

where the average complexity weight of stored pictures is $\bar{W}_{p,i}(j)$, and $\bar{W}_{b,i}(j)$ is the average complexity weight of the unstored pictures. To obtain their values the following expressions are used:

$$\bar{W}_{p,i}(j) = \frac{W_{p,i}(j)}{8} + \frac{7\times\bar{W}_{p,i}(j-1)}{8} \tag{4.71}$$

$$\bar{W}_{b,i}(j) = \frac{W_{b,i}(j)}{8} + \frac{7\times\bar{W}_{b,i}(j-1)}{8} \tag{4.72}$$

$$W_{p,i}(j) = b_p(j)\times QP_{p,i}(j) \tag{4.73}$$

$$W_{b,i}(j) = \frac{b_b(j)\times QP_{b,i}(j)}{1.3636} \tag{4.74}$$

where $b_p(j)$ and $b_b(j)$ are the number of bits generated by encoding the corresponding frame, $QP_{p,i}(j)$ and $QP_{b,i}(j)$ are the corresponding quantisation parameters in the jth frame, and $W_{p,i}(j)$ and $W_{b,i}(j)$ are the weight of the current/previous stored picture and unstored picture. When the number of non-stored pictures between two stored pictures is zero, then Equation (4.70) can be simplified as follows

$$S_i(j+1) = S_i(j) - \frac{S_i(2)}{N_p(i)-1} \tag{4.75}$$

Based on the value of the target buffer, the target bits for the j[th] stored picture in the i[th] GOP is determined based on the target buffer level, Equation (4.70), the frame rate, the available channel bandwidth, and the actual buffer occupancy of (4.70) as follows:

$$T_i'(j) = \frac{R_i(j)}{f} + \gamma\times(S_i(j) - V_i(j)) \tag{4.76}$$

In Equation (4.76), γ is a constant, and its typical value is 0.5 when there is no unstored picture (B frames) and 0.25 otherwise. The remaining bits can be computed as follows:

$$\hat{T}_i(j) = \frac{W_{p,i}(j-1) \times B_i(j)}{W_{p,i}(j-1) \times N_{p,r} + W_{b,i}(j-1) \times N_{b,r}} \tag{4.77}$$

where $N_{p,r}$ and $N_{b,r}$ are the number of the remaining stored pictures and the number of the remaining unstored pictures, respectively. Thus, the target bits can be computed as a weighted combination of $T_i'(j)$ and $\hat{T}_i(j)$:

$$T_i(j) = \beta \times \hat{T}_i(j) + (1-\beta) \times T_i'(j) \tag{4.78}$$

where β is a constant (a typical value is 0.5 when there is no non-stored picture and 0.9 otherwise [169],[170],[339]). Analysing Equation (4.78) it can be observed that tight buffer regulation can be accomplished by selecting small values of β. In the same way, a large value of β should be selected when the goal is to improve video distortion caused by a scene change. To conform with the virtual buffer constrains, the target bits are bounded by

$$\begin{aligned} T_i(j) &= \max\left\{Z_i(j),\ \mathrm{T_i}(j)\right\} \\ T_i(j) &= \min\left\{U_i(j),\ \mathrm{T_i}(j)\right\} \end{aligned} \tag{4.79}$$

where Zi(j) and Ui(j) are determined by

$$Z_i(j) = \begin{cases} B_{i-1}(N_{i-1}) + \dfrac{R_i(j)}{f} & j = 1 \\ Z_i(j-1) + \dfrac{R_i(j)}{f} - b_i(j) & other \end{cases} \tag{4.80}$$

$$U_i(j) = \begin{cases} (B_{i-1}(N_{i-1}) + t_{r,1}(1)) \times \varpi & j = 1 \\ U_i(j-1) + (\dfrac{R_i(j)}{f} - b_i(j)) \times \varpi & other \end{cases} \tag{4.81}$$

where $t_{r,1}(1)$ is the removal time of the first picture from the coded picture buffer and ϖ is a constant with typical value of 0.9 ([169],[170],[339]).

In the next step, quantisation parameter is determined and RDO is performed. After calculating the target bits for the stored picture, the quantisation parameter can be determined according to the following quadratic model:

$$T_i(j) = c_1 \times \frac{\tilde{\sigma}_i(j)}{Q_{step,i}(j)} + c_2 \times \frac{\tilde{\sigma}_i(j)}{Q^2_{step,i}(j)} - m_{h,i}(j) \tag{4.82}$$

where c_1 and c_2 are two coefficients, $m_{h,i}(j)$ is the total number of header bits and motion vector bits, and $\tilde{\sigma}_i(j)$ is the complexity. Typically, the mean absolute difference (MADs) is used as the distortion measure, and it should be the actual MAD of the current stored picture. The MAD of the current stored picture, the value of $\tilde{\sigma}_i(j)$, is predicted with a linear model using the actual MAD of the previous stored picture, $\sigma_i(j-1-L)$ by the following equation

$$\tilde{\sigma}_i(j) = a_1 \times \sigma_i(j-1-L) + a_2 \tag{4.83}$$

where a_1 and a_2 are two coefficients. The initial values of the coefficients is set to 0 and 1 (respectively, a_1 and a_2). After encoding a frame, these parameters are updated by a linear regression method similar to that of MPEG-4 Q2. To update the model coefficients c_1 and c_2, a least square estimation is performed using a set of actual data on the target bits, complexity, and quantisation steps as stated in [23].

The corresponding quantisation parameter $QP_i(j)$ is then determined by using the relationship between the quantisation step and the quantisation parameter of H.264/AVC. To preserve the smoothness of visual quality among consecutive frames, the quantisation parameter $QP_i(j)$ is further adjusted by

$$QP_i(j) = \min\left\{QP_i(j-L-1)+2, \max\left\{QP_i(j-L-1)-2, QP_i(j)\right\}\right\} \tag{4.84}$$

The final quantisation parameter is further bounded to the interval between 0 and 51. RDO is performed for each macroblock, in the current frame, using the quantisation parameter. Consider macroblock s to be encoded. This process can be summarized in the following algorithm for motion estimation and mode decision ([325]):

1. Given the last decoded frames, λ_{MODE}, λ_{MOTION}, and the macroblock quantiser QP.

2. Select intra prediction modes for the *INTRA 4x4* macroblock mode that minimise:

$$J(s,c,IMODE \mid QP,\lambda_{MODE}) = SSD(s,c,IMODE \mid QP) + \lambda_{MODE} \cdot R(s,c,IMODE \mid QP) \tag{4.85}$$

with $IMODE \in \{DC, HOR, VERT, DIAG, DIAG\_RL, DIAG\_LR\}$.

3. Find the best $INTRA16x16$ prediction mode by selecting the mode that generates the minimum SATD.

4. For each 8x8 sub-partition

Perform motion estimation and reference frame selection by minimizing

$$\text{SSD} + \lambda \text{Rate}(\text{MV}, \text{REF})$$

$\text{B} - \text{frames}:$ Choose prediction direction by minimizing

$$\text{SSD} + \lambda \text{Rate}(\text{MV}(\text{PDIR}), \text{REF}(\text{PDIR})) \tag{4.86}$$

Determine the coding mode of the 8x8 sub – partition using the rate – constrained mode decision,

$$\text{SSD} + \lambda \text{Rate}(\text{MV}, \text{REF}, \text{Luma} - \text{Coeff}, \text{block 8x8 mode})$$

5. Perform motion estimation and reference frame selection for 16x16, 16x8, and 8x16 modes by minimizing

$$\begin{aligned} J(REF, m(REF) \mid \lambda_{MOTION}) = SA(T)D(s, c(REF, m(REF))) \\ + \lambda_{MOTION} \cdot (R(m(REF) - p(REF)) + R(REF)) \end{aligned} \tag{4.87}$$

6. B-frames: Determine prediction direction by minimizing

$$\begin{aligned} J(PDIR \mid \lambda_{MOTION}) = SATD(s, c(PDIR, m(PDIR))) \\ + \lambda_{MOTION} \cdot (R(m(PDIR) - p(PDIR)) + R(REF(PDIR))) \end{aligned} \tag{4.88}$$

7. Select the macroblock prediction mode that minimises

$$J(s, c, MODE \mid QP, \lambda_{MODE}) = SSD(s, c, MODE \mid QP) + \lambda_{MODE} \cdot R(s, c, MODE \mid QP) \tag{4.89}$$

given QP and λ_{MODE} when varying $MODE$. $MODE$ indicates a mode out of the set of potential macroblock modes

$$\begin{aligned} MODE &\in \{INTRA4x4, INTRA16x16\} && \text{I} - \text{frame} \\ MODE &\in \begin{Bmatrix} INTRA4x4, INTRA16x16, SKIP, \\ 16x16, 16x8, 8x16, 8x8 \end{Bmatrix} && P - \text{frame} \\ MODE &\in \begin{Bmatrix} INTRA4x4, INTRA16x16, \ DIRECT, \\ 16x16, 16x8, 8x16, 8x8 \end{Bmatrix} && B - \text{frame} \end{aligned} \tag{4.90}$$

It should be noted that the $SKIP$ mode refers to the 16x16 mode where no motion and residual information is encoded. SSD is the sum of the squared differences between the original block s and its reconstruction c given as

$$SSD(s,c,MODE\,|\,QP) = \sum_{x=1,y=1}^{16,16} \left(s_Y[x,y] - c_Y[x,y,MODE\,|\,QP]\right)^2 + \sum_{x=1,y=1}^{8,8} \left(s_U[x,y] - c_U[x,y,MODE\,|\,QP]\right)^2 + \sum_{x=1,y=1}^{8,8} \left(s_V[x,y] - c_V[x,y,MODE\,|\,QP]\right)^2, \quad (4.91)$$

and $R(s,c,MODE\,|\,QP)$ is the number of bits associated with choosing $MODE$ and QP including the bits for the macroblock header, the motion, and all DCT blocks. $c_Y[x,y,MODE\,|\,QP]$ and $s_Y[x,y]$ represent the reconstructed and original luminance values; c_U, c_V and s_U, s_V the corresponding chrominance values. The Lagrangian multiplier λ_{MODE} is given by ([325]):

$$\begin{aligned} \lambda_{MODE\_IP} &= 0.85 \times 2^{(QP-12)/3} && I,P \text{ slices} \\ \lambda_{MODE,B} &= \max\left(2, \min\left(4, (QP-12)/6\right)\right) \times \lambda_{MODE\_IP} && B \text{ slices} \end{aligned} \quad (4.92)$$

If the picture is P or B ([325]) and the SAD is adopted as the criterion, the lambda in motion estimation is give by

$$\lambda_{MOTION} = \sqrt{\lambda_{MODE}} \quad (4.93)$$

When the basic unit is not defined as a frame, an additional basic unit level rate control needs to be added to the process. A basic unit (BU) is defined as a group of continuous MBs. It is used to provide a trade-off between the overall coding efficiency and the bit's variation. A basic unit consists of N_{mbunit} MBs, where N_{mbunit} is a fraction of N_{mbpic}. The total number of basic units in a frame, N_{unit}, is computed by

$$N_{unit} = \frac{N_{mbpic}}{N_{mbunit}} \quad (4.94)$$

If N_{mbunit} is equal to N_{mbpic}, it corresponds to picture level rate control, and if N_{mbunit} is equal to 1, then to macroblock level rate control. The basic unit level rate control is similar to the picture level rate control, including MAD prediction, bit allocation, and quantisation parameter decision in basic unit level ([331]). Further information about this process is available at ([339],[340]).

4.3 Summary

There are two key research issues in the rate control algorithms: the buffer occupancy and the rate distortion model ([282]). A rate control strategy based on buffer occupancy targets to avoid

buffer overflow and underflow. The rate control can be based on the current encoder buffer or on a virtual model of an encoder buffer. In this case, the HRD behavior is often described only at the picture level. To be able to operate at the macroblock level, in some rate control algorithms, such as MPEG-2 TM5, an additional “imaginary” buffer of the encoder is defined. This approach also allows the tuning of the buffer parameter's values, like buffer occupancy, at any time. Rate distortion models allow control of the bit rate according to a specific target bit rate that can define even at the frame level, and models have been used in video coding models.

Video coding standards have their own recommendation on rate control as an informative part based on the work developed during the development phase. These rate control algorithms do not aim to deliver an optimal solution. The adopted rate control algorithms are competitive in R-D performance, with acceptable computational complexities, and are flexible in terms of adaptation capacities regarding different video sources. Thus, it is valuable to review rate control algorithms as they incorporate the progresses obtained in recently developed rate control techniques. Moreover, the level of performance obtainable by these methods serves as a point of comparison for future research and the development of rate control methods. Most rate control algorithms studied in video coding standards are designed for both CBR and VBR applications. Rate control schemes were developed for the simulation models of the different video coding standards, e.g. test model 5 (TM5) for MPEG-2 ([48]), test model near-term (TMN) for H.263 ([53]), verification model (VM) for MPEG-4 ([49]), and joint model (JM) for H.264 ([97]). In this section, the background ideas of these rate control schemes is reviewed, representing the state-of-the-art technology of hybrid video coding.

The MPEG-2 TM5 proposes a rate control algorithm for CBR encoding. The algorithm employs a simple first-order RD model ($R = X/Q$) and a virtual buffer for each type of picture. First, a bit budget is allocated to a GOP. Within the GOP, bits are allocated to frames regarding the available bit budget, the average bit rate and relative complexity of the frame type. The fullness of the virtual buffer is updated after each macroblock has been encoded. Then, the quantisation parameter is determined for the next macroblock. This value is modulated based on buffer occupancy and on a special activity measure. The H.263 TMN8 describes a rate control algorithm composed of two levels: frame and macroblock level. The development of a Lagrangian coder control and parametric choice has led to the creation of the test model TMN-10 ([322],[323]). TMN-10 has two different methods for determining motion vectors and macroblock coding modes: a low complexity mode using a fast block-matching algorithm, and a high-complexity / high-performance mode using a rate-distortion optimisation algorithm ([322],[323]). The MPEG-4 VM8 presents a rate control algorithm based on the quadratic R-Q model. This model is used for the different type of frames. A sliding window mechanism is

introduced to adjust the parameters of the model. The MPEG-4 VM8 algorithm uses a measure of encoding complexity based on MAD that is easier to determine compared with the empirical variance of the H.263 TMN8 model. This algorithm also differs from H.263 TM8 as it aims to achieve a target bit rate over a set of frames. Thus, it can provide a uniform visual quality.

The H.264/AVC JM describes a rate control algorithm that employs a rate-distortion (RD) optimisation technique and is compliant to the H.264/AVC standard HRD. Although the rate control in H.264/AVC JM works well in controlling the rate, there are some aspects that could make rate control more effective and accurate, which are worth studying. The next chapter will focus on this topic.

Chapter 5. Rate Distortion Modeling for H.264/AVC

The main goal of rate control is to provide the proper coding parameters to achieve optimal video quality and compression performance within constricted boundaries such that the buffer does not overflow or underflow, or limited storage capacity or channel bandwidth ([18],[331]).

In Chapter 3, various video compression standards were introduced, with an ever-increasing number of coding parameters. Rate control algorithms adjust these parameters in order to satisfy the requirements of different applications. Chapter 4 introduced various rate control algorithms, developed during the standardisation of recent video coding standards. In general, a rate control algorithm has two steps: resources allocation and computing of the quantisation parameter. The first step can be performed among different video objects (the rate control of multiple video objects), different frames (the rate control of single sequence) or different sequences (the rate control of joint sequences, which is the focus of the present work). In the second step, the coding mode is selected and the quantisation parameter computed, usually based on a R-D (Rate-Distortion) model and RDO (Rate Distortion Optimization) process.

In this chapter, previous research conducted in the field of R-D modelling is examined and R-D functions for the rate control of joint video sequences, appropriate to real video applications, are developed.

5.1 Introduction to Rate Control Optimization

Rate control manages the process of bit allocation within a video sequence and thus the quality of the encoded bitstream. Regarding rate control, encoders can operate at Constant Bit Rate (CBR) or at Variable Bit Rate (VBR). In CBR, the video encoder maintains the average bit rate stable. Likewise, the quality of the video sequence varies due to the changes in scene complexity. VBR reduces the variation in the picture quality by allocating more bits to complex images. A common use of VBR is *Open-Loop Variable Bit Rate* (OL-VBR), where the quantisation scale is fixed for all the images of the video sequence. Another VBR scheme is *Constant Quality - Variable Bit Rate* (CQ-VBR) that aims to maintain the objective video quality unvarying. A generic bit rate control tries to solve the following challenge: given an

input video signal and a desired bit rate, constant or variable, what should be the encoder settings to maintain the picture quality as high and stable as possible? (Figure 5.1)

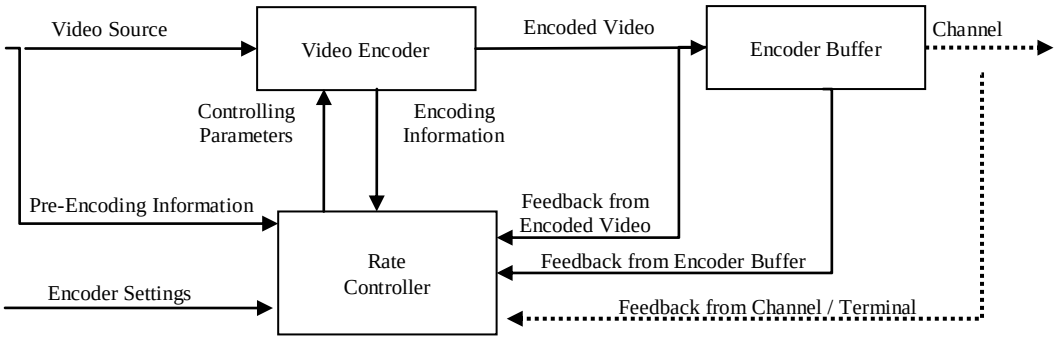


Figure 5.1 – Rate Control in Video Coding System

In MPEG encoding, a quantisation scale controls the trade-off between picture quality and the bit rate. This parameter is used to compute the step size of the uniform quantisers used for the different transform coefficients ([19]). As MPEG does not specify how to control the bit rate, various approaches can be found in the literature ([19],[345],[346]).

One way to approach rate control is to analyse the type of feedback information available: the '*feed forward bit rate control*' and the '*feed backward bit rate control*'. In the first approach, after performing a pre-analysis, the optimum settings are computed. This process will increase the computational complexity and time needed while yielding better results. In the second approach, there is limited knowledge of the sequence complexity. Bits are allocated on a picture basis and spatially uniform distribution throughout the image. Thus, too many bits may be spent at the beginning of the picture while the end of the picture may present a higher degree of complexity. The '*feed backward bit rate control*' is suitable for real time applications and '*feed forward bit rate control*' for applications where the quality is the main goal, and time is not a constraint.

In several rate control algorithms, R-D models are used to select the coding mode and the quantisation parameter is computed. The goal of rate-distortion theory is to determine a lower bound on the rate so that a known source is characterised with a given fidelity, or conversely ([347],[348]). Thus, rate-distortion theory aims to determine what is the fewest numbers of bits necessary to represent a source, subject to a fidelity criterion. One can differentiate between two classes of bounds, models based on Shannon Rate-Distortion theory ([347],[348],[349],[350]) and those derived from high-rate quantisation theory ([347],[351]). The first presents asymptotic results as the sources are encoded with longer and longer blocks. The second presumes fixed input block sizes. Performance is estimated as the encoding rate becomes arbitrarily large.

These two approaches are complementary ([347],[350]), and converge to the same lower bound $D \sim e^{-\alpha R}$ when the input block size increases to infinity, as shown in [351]. More information and an assessment between these two approaches can be found in [352]. There are two main concerns with the R-D theoretical approach to determine bounds: complexity (is it possible to implement the algorithm to determine the bound? How much delay or computational resources are needed?), and model mismatch ([350]).

Two approaches to solve the optimal bit allocation problem are the Lagrange's optimization ([15],[16]) and the Dynamic Programming (DP) ([17]). The optimal bit allocation was first addressed in [353] where the Lagrange multiplier approach for R-D analysis in transform coding was used. Further improvements have been reported in [354] for source quantisation and coding. Lagrangian method has been used to solve the constrained optimization problem in a continuous setting ([355]). In the case of the discrete optimization problem, it has also been used to relax the original constrained problem into an easier unconstrained problem. The optimal solution can be obtained by iteratively searching for the Lagrangian multiplier that generates the closest results that satisfy the constraint. Lagrange's multiplier method suffers from problems, such as having negative bits and real numbers ([356]), and the computational complexity is very high because it needs to determine R-D characteristics of current and future video frames. DP is a robust tool in solving optimization problems. DP has been used to achieve the minimum overall distortion through a tree or trellis with known quantisers and their R-D characteristics ([15],[18],[357]). One interesting application of DP is when a set of optimization problems can be broken down into sub-problems. Then, each sub-problem should contain three states (a past, the current and a future state). DP will efficiently produce a globally optimal solution if the future problem solution does not depend on the past problem solution ([355]).

Rate control algorithms can also be classified according to the criteria used for determining how to allocate resources and to compute the quantisation parameter: buffer-based approaches, analytic modelling approaches, and the operational rate-distortion (R-D) modelling approach ([358]). In the first approach, estimation is based on the buffer occupancy, local scene activity, and bit count of the previous frame. It is used in real-time and low-complexity applications, but if a scene change takes place, then quality degradation occurs as decisions are made based on past information. In the second approach, estimation is based on a set of "rate-quantisation" and "distortion-quantisation" functions. These functions are obtained using analytic models with parameters computed from coding statistics generated by the recoding of input video. To increase the model's accuracy, different techniques for data fitting can be used at the expense of computational complexity. Therefore, this approach is generally not used in real-time applications. In the last approach, allocation is based on prediction of the signal statistics from

already processed pictures. In this case, computational complexity is reduced at the cost of prediction accuracy.

RD modelling has also been divided into empirical approaches, analytical approaches and operational R-D model ([24],[25],[359],[360]). The first approach, as explained above, estimates the R-D model by mathematical processing of the statistical information, such as interpolating between (R-D) samples of a given encoder ([24],[25],[361],[362],[363]). The second approach generates R-D models based on rate-distortion or quantisation theory with certain assumptions of source statistical properties ([21],[22],[23],[360]).

Usually, they are based on the assumption that video source statistics are stationary. In this case, video source statistics correspond to some form of a probability model such as Gaussian distribution ([22],[364]), generalized Gaussian and Laplacian distributions ([21],[23]), the p domain analysis ([365],[366]), and centralized Cauchy's density ([367]). The third approach will be described with more detail. An R-D model should be simple enough so that it achieves good performance at reasonable cost, but presents a sufficient degree of complexity that enables it to describe the principal characteristics of the source. To obtain a practical solution, firstly the coding algorithm and a set of test sequences are selected. Then, one must search the best operating points for that particular system. When both R and D are quantified, the maximum value of distortion or fidelity can be varied, and the corresponding bound R is determined. Following this procedure a pair (R,D) is obtained. This corresponds to a mode of operation of the encoder, for a particular set of parameters and for a given set of test data. These pairs may be represented by a curve in the R–D space, defined as the operational rate-distortion (ORD) curve.

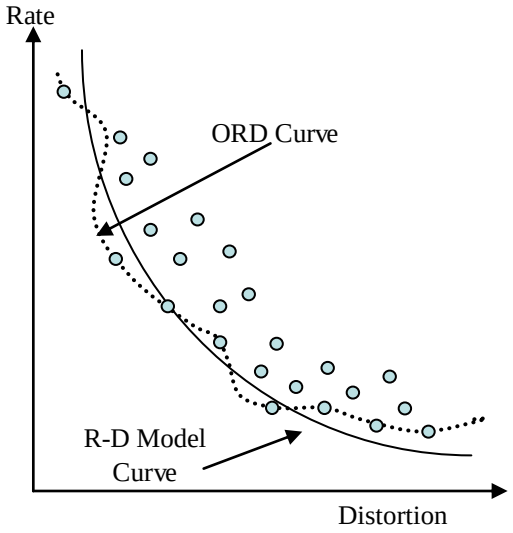


Figure 5.2 – Operational rate-distortion and rate-distortion model curves

The ORD function presents the convex curve of the specific compression scheme such that the optimal solution of rate control, i.e., optimal quantiser achieving minimum distortion at a given bit rate, can be obtained ([356]) (Figure 5.2). The set of coding parameters that result in the ORD function can be formally defined as

$$Q_{ORDF} = \{q : q \in Q, R(q) \geq R(p) \Rightarrow D(q) \geq D(p), \forall q \in Q\} \quad (5.1)$$

where $R(q)$ and $D(q)$ are the rate and distortion generated by a particular coding mode with the quantiser q. The bound obtained this way allows differentiating between the best feasible operating points and those that are suboptimal or unattainable ([350]). Efficiency problems in many practical video coding applications may occur due to the high computational complexity in this approach ([368]). Thus, simpler methods, such as the so-called model-based R-D functions, are frequently adopted ([23],[24],[32],[369],[370]). In some cases, this alternative approach may experience considerable estimation errors, as shown in Figure 5.2 where the solid curve represents an approximation of the ORD curve by a model-based R-D function.

In practical video coding applications, the rate and the coding distortion depend on the quantisation step-size. The rate or distortion versus quantisation parameter (Q) curve can be produced by encoding for all the possible quantisers to obtain the bit rate and the quantisation error. In order to know how to select a quantisation parameter under a specific constraint, e.g. the target bit-budget or distortion, it is important to model or estimate the rate in terms of the quantisation parameter, namely rate-quantisation (R-Q) functions. Together with the distortion-quantisation (D-Q) function, R-Q function characterises the rate-distortion (R-D) behaviour of video encoding, which is the key to obtaining an optimum bit allocation. Many R-Q and D-Q functions have been reported in previous studies ([18],[19],[20],[21],[22],[23],[24],[25],[26], [27],[28]). Some of these schemes were adopted in standard-compliant video coders, as described in Chapter 4. Some rate control schemes incorporate spatio-temporal correlations to improve the accuracy of R-D models, by using statistical regression analysis for dynamical model parameters' updates. Representative of this approach is the MPEG-4 VM8 and H.264/AVC JM, where model parameters are updated by the linear regression method from previous coded parameters. The H.264/AVC rate-control uses the Lagrange approach to determine the motion vectors and mode decisions ([169],[170],[323]).

Over the past few years, extensive research has been conducted on optimizing a video encoder with R-D considerations including mode decision ([296],[371]), motion estimation ([372],[373]), optimal bit allocation and rate control in video coding field, etc ([15],[18],[287], [298],[374],[375],[376],[377],[378],[379],[380],[381]). In this work, H.264/AVC JM is the platform where the proposed techniques are integrated. Thus, its performance is the benchmark

against which the proposed techniques are compared. The next section will discuss several ways to optimize the H.264/AVC algorithm.

5.2 Rate Control Optimization in H.264/AVC JM Model

The JM rate control algorithm consists of three levels, including: GOP level, picture level and an optional basic unit-level rate control when the basic unit is not defined as a frame. In the GOP level rate control, the number of available bits for the uncompressed pictures in the GOP is computed based on the picture type, the total number of pictures in the GOP, the instant bit rate, the occupancy of the virtual buffer, and the actual generated bits by the encoded pictures. The picture level rate control is performed in two stages: pre-encoding and post-encoding. In the pre-encoding stage, the quantisation parameter is computed, and in the post-encoding stage, parameters are updated. Every macroblock, in the same basic unit, share a common quantisation parameter. To determine its value, a quadratic R-D model is used. A linear model was selected to estimate the value of MAD of the current stored picture by using the MAD value of the previous stored pictures. The basic unit layer rate control adjusts the values of the quantisation parameters of the basic units in a frame, so that the sum of generated bits does not exceed the frame target bits.

The H.264/AVC JM model experiences other performance difficulties because MAD is predicted through a linear model in contrast to MPEG-4 VM8 that uses the actual value of MAD. Several alternatives have been proposed to overcome this problem. Authors in [382],[383] proposed an alternative that is the ratio between the predicted MAD of the current frame and the MAD average value of all previously encoded P frames in the GOP. Based on this work, a normalised MAD has been proposed ([384],[385]). A normalised MAD is the ratio of the enhanced predicted MAD for the current frame i, MADEPi, to the average value of MAD for all the previously encoded P frames in a GOP ([384],[385]). Zhao Min in [386] proposes an improvement of the MAD prediction model by using a model based on the variance of two models: a temporal MAD model and a spatial MAD model. Comparing variances of these models allows the selection of which model should be chosen when predicting the next basic unit. Do-Kyoung Kwon in [387] suggests replacing MAD by the SATD of coded 4x4 blocks. This proposal aims to provide a better characterisation of the complexity of the residual signal. Furthermore, Do-Kyoung Kwon in [387] suggests dropping the second-order term in Equation (4.82) resulting in a simplified source model.

$$T_i(j) = \alpha \times \frac{SATD_i}{Q_{step,i}(j)} \tag{5.2}$$

Simultaneously, the use of look-ahead processing techniques allow exploring the complexity of the frames to be encoded and gather their R-D statistics for joint video coding. Dong et al. ([388]) found that the linear R-D model has better predictability and is faster to estimate than the quadratic R-D model, especially for timing-sensitive applications. With good predictability, the linear R-D model may have a better rate-distortion performance than the quadratic R-D model. Dong improves the method of updating the model parameters of X as follows

$$X = \frac{\sum_k \frac{A_k \times MAD_k}{Q_{step,k}}}{\sum_i \left(\frac{MAD_k}{Q_{step,k}} \right)^2} \tag{5.3}$$

where MADk, Ak, and Qstep,k denote the MAD, actual texture bits, and Qstep for a basic unit k in the sliding window, respectively.

To improve coding efficiency, a rate-distortion (RD) optimization technique is used. An exhaustive search is performed to find the best coding mode and the best motion vector that optimize the rate-distortion among all possible coding modes and among all possible motion vectors, respectively. The best coding mode decision for an Intra macroblock is obtained after assessing 22 possible modes (nine modes for the 4x4 block, nine modes for the 8x8 block and four modes for the 16x16 block). In the case of a P macroblock, the search space includes the Intra modes and the additional Inter prediction modes. For a macroblock in a B-slice, an encoder has a great number of coding options, such as all the intra modes, all the inter partition sizes, all the possible motion vectors, all the available reference frames and the choice of one-directional or bi-predicted motion compensation. Performing a complete search through all possible modes and motion vectors to find the best combination of mode and prediction type is a highly computationally intensive task. In a real encoding scenario, it requires the following steps. First, the predicted block is determined, and then the residual block. Afterwards, the integer and scale transform is applied to the residual block. The next step is to quantise the transformed residual block followed by entropy coding of the quantised and transformed residual block. The inverse process is then performed: inverse quantize the quantized and transformed residual block, inverse integer and scale transform the dequantized block, compute the reconstructed image block. Finally, SSD between the original block and reconstructed block is determined, and the value of the RD cost function is calculated.

The computational cost makes it very difficult to perform RDO in real-time applications. This process is more highly complex than in previous MPEG standards. For example, Equation (5.4) displays the different ways that a macroblock belonging to a P-slice can be encoded ([298]).

$$\begin{array}{ll} MPEG-2 & \{INTRA, SKIP, INTER16x16\} \\ H.263/MPEG-4 & \{INTRA, SKIP, INTER16x16, INTER8x8\} \\ H.264 & \left\{\begin{array}{l} INTRA4x4, INTRA16x16, SKIP, \\ INTER16x16, INTER16x8, \\ INTER8x16, INTER8x8 \end{array}\right\} \end{array} \quad (5.4)$$

It should be noted that even if the names are the same in some cases, the modes differ regarding the standard. Notice that when a P-slice macroblock is coded in INTER-8×8 mode, then H.264/AVC specifies a set of sub-macroblock types for each 8×8 sub-macroblock: INTER-8×8, INTER-8×4, INTER-4×8, and INTER-4×4.

Decreasing the computation cost for RDO techniques has become one of the principal research tasks in video compression now. Several proposals that explore fast algorithm in motion estimations, intra mode predictions and inter mode predictions for H.264/AVC video coding have been made ([389]). One of the popular proposed approaches is to reduce the number of mode decisions. Meng at al. ([390]) compute only a partial cost for down-sampled pixels instead of a 4x4 block. Chun-Ling Yang et al. ([391]) present a fast 16x16 intra mode decision based on macroblock properties. Feng Pan et al. ([392]) describe an approach based on the use of the edge map of the whole frame in order to obtain the best interpolation direction. This proposal demands high computational resources as every pixel in the frame needs to be evaluated. To reduce the complexity of intra mode decision, SAD and SATD have also been proposed as cost functions, in JM 6.1 (Equation (5.5) and Equation (5.6)).

$$J_{SAD} = \begin{cases} SAD(S,P) + \lambda_1 \bullet 4K & if \text{ intra } 4x4 \text{ mode} \\ SAD(S,P) & otherwise \end{cases} \quad (5.5)$$

$$J_{SATD} = \begin{cases} SATD(S,P) + \lambda_1 \bullet 4K & if \text{ intra } 4x4 \text{ mode} \\ SATD(S,P) & otherwise \end{cases} \quad (5.6)$$

where SAD(S,P) is the sum of absolute differences between the original block S and the predicted block P, SADT(S,P) is the sum of absolute Hadamard transformed differences between the original block S and the predicted block P. The value of λ_1 is also approximate by an exponential function of the QP, and the K is equal to 0 for the probable mode and 1 for the other modes. These functions manage to reduce the computation complexity at a cost of a degradation of performance. Notice that the Lagrangian optimization technique can be expressed as

$$\min\{D\} \quad where \;\; J = D + \lambda \cdot R \quad (5.7)$$

where J is named the Rate Distortion (RD) Cost, D is the distortion, R is the number of bits per pixels, and λ is known as the Lagrange multiplier. If the R-D curve is convex, and both rate and distortion are differentiable everywhere, the minimal value of J, for an access unit is obtained by setting its derivative to zero as follows

$$\frac{\partial J}{\partial R} = \frac{\partial D}{\partial R} + \lambda = 0 \tag{5.8}$$

Solving Equation (5.8), the value of λ is computed by

$$\lambda = -\frac{\partial D}{\partial R} \tag{5.9}$$

In JM, the rate-distortion function is derived as follows ([323]):

$$R(D) = a \log\left(\frac{b}{D}\right) \tag{5.10}$$

with the parameters a and b whose values depend on the content. Using a uniform distribution to estimate the source probability within each quantization interval, one can determine the Lagrange multiplier

$$\lambda = -\frac{\partial D}{\partial R} = c \times 2^{(QP-12)/3} \tag{5.11}$$

where c is a constant equal to 0.85 ([323]). This value was proposed based on empirical results and typical RD models ([27],[323]). In these papers, it is also proposed that λ is a function of QP only, and as a result is independent of the content properties. Although this hypothesis simplifies the problem, in some cases, it may not yield the optimal λ as macroblocks have different perceptual relevance. This has inspired research on how to adapt λ according to video content, at sequence, frame and macroblock levels.

To improve the R-D performance in both mode decision and rate control, the sum of absolute integer transform difference (SAITD) have been proposed by Tseng ([393]). The main disadvantage of this approach is the bit estimation method whose performance decreases. Mohammed in [389] proposes a method for estimation rate cost of the intra and inter mode decision, based on the properties of CAVLC and observations of VLC tables.

Shuijiong Wu et al. describe, in [394],[395],[396], a multi-stage rate control scheme for R-D optimized H.264/AVC encoders under CBR video transmission. A frequency-domain parameter, designated by mean-absolute-transform-difference (MATD), is introduced to represent frame and macroblock (MB) residual complexity (Equation (5.12)),

$$\begin{cases} SATD\_4\times4 = \sum_{m=1}^{4\times4}\left|T\left\{C_m(x,y)-P_m(x,y)\right\}\right| \\ MATD\_mb = \dfrac{\sum_{n=0}^{15} SATD\_4\times4_n}{16\times16}\times\eta \\ MATD\_f = \dfrac{\sum_{n=1}^{N_{mb}} MATD\_mb_n}{N_{mb}} \end{cases} \tag{5.12}$$

with $SATD\_4\times4$ the value of SATD of each 4×4 block using Hadamard transform, $MATD\_mb$ and $MATD\_f$ the MATD values of each macroblock and frame, respectively, N_{mb} the number of macroblocks in one frame, and η is an adaptive coefficient depending on the sequence features. MAD is replaced by MATD in the estimation of complexity. This change occurs both at frame and macroblock layer because of its slightly better performance in the source rate model. In [397], Chen et al. propose an adaptive λ estimation algorithm using a R-D model in ρ domain, where ρ is defined as the percentage of zero coefficients among quantised transform residuals ([375]). Chen uses R and D models defined as follows

$$\begin{aligned} R(\rho) &= \theta\cdot(1-\rho) \\ D(\rho) &= \sigma^2\cdot e^{-\alpha(1-\rho)} \end{aligned} \tag{5.13}$$

where θ and α are coding constants, σ^2 is the variance of transformed residuals. Thus, the value of λ is

$$\lambda_\rho = \beta\cdot\left(\ln\left(\frac{\sigma^2}{D}\right)+\delta\right)\cdot\frac{D}{R} \tag{5.14}$$

where β and δ are both coding constants. In [398], Xiang et al. presents a Laplace distribution based on a Lagrangian RDO algorithm for one-pass coding. Accurate rate and distortion models are first obtained based on Laplace's distribution of transformed residuals. The Lagrange multiplier is subsequently computed from the models. A special designed escape method was developed when the statistical properties of input sequences cannot be well captured by the models. This approach presents good results, but the formula for λ is rather complex and may not be easy to implement in practice. Xiang el al. also propose, in [399], a one-pass multi-layer RDO algorithm for quality scalable video coding. The goal is to improve the coding efficiency while keeping reasonable computational complexity. Xiang proposes to estimate the impact of the base layer on the enhancement layer after real coding instead of computing the value. This method avoids the heavy computational multi-pass process.

Wang and Yang, in [400], and Jiang and Ling in [401] proposed estimating λ at the macroblock level instead of using the same value for all macroblocks in an access unit. Following this concept, Zhang et al. in [402] introduced the Context Adaptive Lagrange Multiplier (CALM) selection method. En-hui Yang and Xiang Yu proposed a Soft Decision Quantization (SDQ) algorithm based on the CABAC method ([403],[404],[405]). The SDQ algorithm is used in conjunction with the general RDO framework to jointly design motion prediction and residual coding for the H.264/AVC main profile coding. The approach presents a high level of computational complexity because it uses a full Trellis. It is directed at off-line applications such as video delivery. In [406],[407], Marta K. et al. proposed a rate-distortion optimized quantization (RDOQ) scheme that aims to estimate a solution for finding the quantised coefficients that minimize the RD Cost function for each transform block. The process consists of two steps: a Trellis-based optimization process of the quantization operation for transform coefficients, and quantizing and coding a block with multiple quantizer step-sizes. The method works with both entropy coding methods employed in H.264/AVC, and includes a fast algorithm for macroblock level QP decision. Recently, Fu-Chuang Chen and Yi-Pin Hsu, proposed in [408] a joint RDO with a novel analytically derived equation for the prediction of distortion and a new method for the optimization of QP selection. The distortion prediction method avoids the traditional linear regression, reduces prediction error and decreases computation costs. Thus, it can be implemented for real-time applications. To further improve the performance of block-based transform coding, Xin Zhao et al., in [409], propose the design of rate-distortion optimized transform (RDOT) and a fast RDOT, which contributes to both intra frame and inter frame coding. In RODT, the transform is implemented with multiple transform basis functions (TBF) candidates. These functions are obtained from off-line training. The encoder is thus capable of choosing the optimal set of TBF, when coding each residual block, in terms of R-D performance, and obtaining better energy compaction in the transform domain. The proposal was successfully adopted into the key technical area (KTA) software ([410]).

However, all these methods ignored the perceptual aspect of the RDO method. A recent line of work has focused RDO research towards the distortion metric and its correlation with HVS. Although SSD, SAD; SATD, have been extensively adopted as distortion metrics because of their simplicity, they do not correlate well with HVS. Thus, they cannot measure the pictures' perceptual distortions very well. As presented in Chapter 2, several perceptual image/video quality assessment metrics have been proposed ([64],[109],[112],[128],[163],[164],[184],[185], [186],[187],[188],[189],[190],[193],[194],[197],[198],[199],[200],[201],[202]). Recent studies show progress in HVS-based video coding methodologies.

In [411],[412], Yang et al. explore a perceptually-adaptive video coding (PVC) scheme by incorporating a derived image-domain JND profile into the motion prediction loop, and the filtering process on the motion compensated residues before DCT coding. Perceptual coding quality, measured by peak signal-to-perceptual noise ratio (PSPNR) is improved. In [413], Zhongjie Zhu et al. propose to improve JM rate control algorithm, modifying the quadratic RD model and incorporating into the rate control algorithm some of the main characteristics of the HVS such as contrast sensitivity, multichannel theory, and masking effect. Thus, the QP for each macroblock is adjusted according to its visual sensitivity. The quadratic RD model is improved based on both empirical observations and theoretical analysis and adapted to be

$$R = \frac{a}{\sqrt{Q}} + \frac{b}{Q^2} + c, \tag{5.15}$$

where a, b, and c are model parameters, and they can be computed based on the linear regression method. The method still uses the MAD prediction method and presents a higher computational load compared with the original algorithm. A foveated JND (FJND) model has been proposed by Chen and Guillemot, in [414], in which the spatial and temporal JND models are enhanced by taking into account the foveation properties of the HVS. Given that perceptual acuity increases with decreased distance, the visibility threshold of the pixel of the image increases when the distance between the pixel and the fixation point increases. The FJND model is used for macroblock quantization adjustment in an H.264/AVC encoder. For each macroblock, the QP is optimized based on its FJND information. The Lagrange multiplier is determined to minimise the noticeable distortion of the macroblock. Z. Li et al. present a visual attention-based bit allocation strategy for video compression ([415]). Four different categories of subjective quality-based video coding methods, according to the way of obtaining attention regions, are identified: human-machine interaction methods, machine vision algorithms, methods based on knowledge of human psychophysics, and computational neuroscience models. Z. Li's proposal uses a neurobiological model of visual attention that automatically predicts high saliency regions in unconstrained input frames to generate an attention map (usually called saliency map). The saliency map contains information regarding the location and intensity of relevant parts of the image, allowing resources to be allocated in a non-uniform manner. Results were assessed using an eye-tracking-weighted PSNR (EWPSNR) measure of subjective quality. Sequences were encoded with the JM9.8 encoder, intra period = 30, Hadamard transform, UVLC, no fast motion estimation, no B frames, high complexity RDO, no restriction on search range, four baseline QPs (24, 28, 32 and 36), and the bit rates range from 260 kbps to 10Mbps. Fifty video clips (1920x1080) were collected for this experience, and all the raw material is available for download. The majority of the encoded video clips achieved

better subjective quality compared with standard encoding. Nevertheless, the proposed attention prediction model purely depends on the bottom-up low-level features. Naccari and Pereira, in [416], have proposed an H.264/AVC-based PVC integrating a decoder side ST-JND (Spatial Temporal) model to perceptually modulate the quantization steps for each DCT coefficient. The model allows rate allocation to be perceptually performed at the finest level of granularity and to avoid extra associated rates to code the varying quantization steps.

As HVS is very complex, it is hard to achieve perfect perceptually reliable results during rate control and video coding ([413],[417]). Thus, because of its performance and simplicity, SSIM has been adopted in several RDO proposals. In general, expression (5.7) is rewritten as follows

$$\min\{J\} \quad where \;\; J=(1-SSIM)+\lambda_{SSIM}\cdot R \tag{5.16}$$

The Lagrange multiplier can be obtained by determining the derivative of J with respect to R and set the result to zero (Equation (5.17)).

$$\frac{\partial J}{\partial R}=\frac{\partial\left((1-SSIM)+\lambda_{SSIM}\cdot R\right)}{\partial R}=-\frac{\partial(SSIM)}{\partial R}+\lambda_{SSIM}=0 \tag{5.17}$$

The value of the Lagrange parameter is then obtained by solving the equation (Equation (5.18)).

$$\lambda_{SSIM}=\frac{\partial(SSIM)}{\partial R}=\frac{\partial(SSIM)/\partial Q}{\partial R/\partial Q} \tag{5.18}$$

Mai et al. in [418] proposed the use of the SSIM index as a distortion metric for an H.264/AVC I-frame encoder. A new Lagrangian multiplier is proposed

$$\lambda_{Mode}=1.11\times 2^{(QP-60)/5} \tag{5.19}$$

where QP denotes the quantization parameter. Results show a small bit rate and perceived quality reduction, with a reduced increase in computation complexity. Babu et al., in [419], implemented using the same Lagrange multiplier. Simulations were performed integrating the algorithm in JM 92 software, QP (10, 20 and 30), and using the first 30 frames of several video sequences. A bit rate reduction of between 0.7% and 3% was obtained. Chun et al. in [420] proposed an Improved Rate-Distortion Optimization method based on SSIM (IRDO-SSIM) for the RDO Inter mode selection process. Motion estimation is performed using SAD to reduce the computational complexity, and then RD Cost is computed using SSIM. A new Lagrange multiplier is derived as follows

$$\lambda_{MODE}=2.39\times e^{(QP+11.804)/6.8652} \tag{5.20}$$

Because SAD is used in the motion estimation process, the gain from the RD performance is limited for high- or low-motion complexity sequences. A recent approach is to develop R-D models to characterize the relationship between rate and distortion in terms of the SSIM index ([421],[422],[423],[424]). A new R-D model which employs the SSIM index as the quality metric can be used in RDO and an optimum bit allocation and rate control scheme developed for video coding. One solution is to model using a power function based upon the MSE-RD curve ([423]). The RD curve can thus be represented by

$$D_{SSIM}(R) = \alpha' \times R^{\beta'} \tag{5.21}$$

where $D_{SSIM} = 1 - SSIM$, and α' and β' are content-dependent parameters ([422],[423]). With the parameters, and a R-D point p of the former encoded picture as a prediction

$$\lambda_{SSIM} = -\frac{D_v - D_h}{R_v - R_h} \tag{5.22}$$

where (Rv, Dv) and (Rh, Dh), respectively, are the vertical and horizontal projections of p to the MSE-based R-D curve. The algorithm was implemented in JM 15.1 with the following simulation conditions for Intra evaluation: Baseline Profile, all frames coded as intra, CAVLC, RDO enabled, QP (24, 28, 32, and 36). Results are presented for 12 video test sequences (five CIF, three D1, two 720p, two 1080p). Subjective evaluation were performed using the DSCQS method. This method has outperformed the JM model. Simulations were also conducted for Inter macroblock mode decision. The simulation conditions were as follows: High Profile, GOP pattern IPPP, CABAC, RDO enabled, QP (16, 20, 24, 28, 32, and 36). Results are presented for 24 video test sequences (sixteen CIF, three D1, five 720p). Preliminary results show that for P frames, at the same SSIM index, an average value of 10% bit rate reduction was obtained.

Nevertheless, for B frames, the R-D characteristic varied significantly between frames and the model showed prediction difficulties. In [424], Tao-Sheng et al. a RD model using SSIM as the quality metric is used, as follows:

$$D(R) = \alpha \times e^{-\beta R} \tag{5.23}$$

This model is used for RDO. For QP determination, the authors have adopted the quadratic R-Q model employed in the JM model. The Lagrange multiplier follows the slope scheme defined in [422],[423].

$$\lambda_{SSIM} = -\frac{\tilde{d}_f - \alpha \times e^{-\beta \tilde{r}_f}}{\tilde{r}_f - \frac{1}{\beta}\ln\left(\frac{\alpha}{\tilde{d}_f}\right)} \tag{5.24}$$

with α and β the model parameters, and the RD data $\left(\tilde{r}_f, \tilde{d}_f\right)$ of the previous encoded frame. The Lagrange multiplier is determined frame by frame and is used for macroblock mode decisions. All the experiments use the JM reference software as the baseline, and the results are compared with the MSE based RDO adopted in the JM reference software and also with the perceptual-based RDO described in [422],[423]. The proposals were integrated in JM15.1 with only the rate control module changed. The experimental configurations were the following: High profile, the first 150 frames of 13 video sequences (eleven CIF and two D1), RDO enabled, search range ±32, fast motion estimation EPZS on, CABAC on, GOP pattern IPPP, 11 macroblocks per basic unit for CIF sequences and 15 for D1 sequences, a key frame every 30 frames. The SSIM gain is positive while the PSNR varies between positive and negative values. This result is normal as the encoders do not optimize decision mode against PSNR. The level of computational complexity is reported to increase.

In a very recent paper (not published at the time of writing), Shiqi Wang et al. proposed an RDO scheme based on a novel reduced-reference statistical SSIM (RR-SSIM) ([417],[425]). The idea is to use an assessment method that requires only a set of RR features extracted from the reference frame for SSIM estimation. The RR-SSIM metric was first proposed in [426] and it has been reported to achieve high SSIM estimation accuracy. Nevertheless, it is based on a multi-scale multi-orientation Divisive Normalization Transform (DNT), which implies high computational complexity. The authors proposed an adaptation of the approach presented in [426] by instead extracting features from the DCT coefficients ([427]). Channappayya et al. ([427]) introduced the FR DCT domain SSIM index as follows.

$$SSIM\left(x,y\right) = \left\{1 - \frac{\left(X\left(0\right) - Y\left(0\right)\right)^2}{X\left(0\right)^2 + Y\left(0\right)^2 + N \bullet C_1}\right\} + \left\{1 - \frac{\sum_{k=1}^{N-1}\left(X\left(k\right) - Y\left(k\right)\right)^2}{\sum_{k=1}^{N-1}\left(X\left(k\right)^2 + Y\left(k\right)^2\right) + N \bullet C_2}\right\} \tag{5.25}$$

where X(k) and Y (k) represent the DCT coefficients for the input signals x and y, respectively. The new RR-SSIM is thus defined as

$$M_{RR} = \left\{1 - \frac{D_0}{\sigma_0^2 + C_1}\right\} \times \left\{1 - \frac{1}{N-1}\sum_{i=1}^{N-1}\left(\frac{D_i}{\sigma_i^2 + C_2}\right)\right\} \tag{5.26}$$

where σi is the standard deviation of the ith subband and N is the block size ([417]). Di represents the MSE between the original and distorted frames in the ith subband, and is calculated as follows

$$D_i = \int_{-(Q-\gamma Q)}^{(Q-\gamma Q)} x_i^2 f_{Lap}(x_i)\,dx_i + 2\sum_{n=1}^{\infty} \int_{n(Q-\gamma Q)}^{(n+1)(Q-\gamma Q)} (x_i - nQ)^2 f_{Lap}(x_i)\,dx_i \tag{5.27}$$

where γ is the rounding offset in the quantization. A new rate model that combines the side information with the entropy of the transformed residuals is proposed and reported to present good results with the exception being at low bit rates. At the macroblock level, the Lagrange parameter is further adjusted based on a spatiotemporal weighting factor. Simulations were performed by integrating the mode selection scheme in JM15.1 with the following coding configurations ([420]): all available inter and intra modes enabled, five references frames, one I frame followed by 99 inter frames, high complexity RDO, the fixed QP are set from 28 to 40, and two GOP structures IPP and IBP. Rate reduction is 14% for IPP GOP pattern and 8% for IBP GOP pattern. The reason for this difference could be that the parameter estimation scheme is not very accurate for this type of GOP pattern. The lower results with B frames need further research.

In the majority of the current and emerging practical real-world visual communication environments, FR methods have a limited application as reference signals are not accessible at the receiver side (or in some cases not at all) ([428]). Thus, a solution based on the use of RR and NR image quality assessment metrics that operate with little or no reference signal information at all; is rather interesting. VQEG experience shows that progress in this field has been slow. RR metrics based on the SSIM index ([429],[430]) or NR metrics such as NORM (NO-Reference Video Quality Monitoring) ([431],[432]) are some of the examples of current proposals for Image Quality Assessment. The integration of these metrics within a video encoder is still a difficult challenge. In summary, to optimize a video encoder, the rate-distortion optimization techniques play a very important role. R-D models are functions that predict the expected distortion at a given bit rate. This is very important for joint video coding applications that attempt to optimize quality, e.g. minimize distortion, in environments where the channel conditions vary dynamically or the number of broadcast programmes varies through time. Most of the work presented focuses on the frame and macroblock layer. For the joint coding of video sequences it is important to compare different R-D models, using traditionally fidelity criterion and perceptual image quality metrics.

5.3 Rate-Distortion Modelling

In this section, several R-D models will be presented and evaluated. The next section will focus on the bit rate variability as a function of video quality ([433],[434]). This type of analysis is characteristic of a communication network perspective. The goal is to find an R-D model that can be use to support the allocating of video bandwidth when different video programmes are jointly encoded.

5.3.1 *Source Materials and Test "Methodology" Configurations*

This section will describe the source material used in all the simulations conducted in Chapter 5 and Chapter 6. Next, brief comments will be presented on how GOP structures can be "produced" in H.264/AVC, quality metrics and encoder configurations.

5.3.1.1 Test Video Sequences

Selecting a representative set of video sequences is a crucial step in evaluating and analysing the performance of R-D models. A homogeneous set of video sequences may generate biased comparison results, because some models may perform especially well under certain sequences. Two key features are used to characterise video sequences: spatial complexity and temporal complexity

a) Akiyo b) Coastguard c) Deadline d) Flower Garden

e) Football f) Foreman g) Hall h) Mobile & Calendar

i) News j) Paris k) Silence l) Mother & Daughter

Figure 5.3 – Video Test Sequences

Usually, the first feature is measured by averaging all neighbourhood differences in the same frame while the second by averaging neighbourhood differences between adjacent frames ([435]). The set of test video sequences is composed of twelve CIF video sequences, frame rate of 30 pictures per second, with a duration of 10 seconds (Figure 5.3) ([436]). The set of video sequences ranges from sequences with low complexity (low spatial and temporal complexity values) to sequences with higher complexity values. It follows a brief description of the sequences.

In seven video sequences, the position of the camera is fixed: Akiyo (aki), Deadline (dea), Hall (hal), Mother and Daughter (mad), News (new), Paris (par) and Silence (sil). In the Akiyo sequence, the camera is centered on a human subject with a synthetic background (a female anchor reading the news). The movements are very limited, mainly head movements in front of a fixed camera. In the Deadline, Mother and Daughter and Paris sequences the camera is still fixed but there is more movement of the bodies and heads. This is typical videoconferencing content. In the News sequence two reporters, a male and a female anchor, read the news in front of a fixed camera in a newsroom while in the background, two dancers execute movements. Hall's sequence is an example of a video supervision, with a stationary camera and two moving persons: one person entering from the left with a briefcase and then leaving the hall. In the middle of the sequence a second person enters the hall from the right and then grabs a monitor. In the Silence sequence, one can observe a fast-moving subject executing deaf gesture language.

The Foreman sequence (for) contains the head of a person talking and geometric shapes. Fast camera movement and content motion with a pan to a construction site at the end characterise this sequence. The main characteristics of the Flower Garden sequence (flg) is the slow and steady camera panning over the landscape; the spatial and the colour detail. Coastguard sequence (cgd) was shot as pan movements. The camera follows the movement of two boats (the first starting from right and moving to the left and the second boat moving in the opposite direction). The Mobile and Calendar (mcl) sequence is characterised by the slow panning and zooming of the camera, complex motion, high spatial and colour detail. The Football sequence (fot) is also characterised by fast and complex motion movements of the camera and the level of detail of the scene. This is a very diverse set of video sequences in terms of spatial and temporal complexity.

5.3.1.2 GOP Structure

To study the performance of a rate control algorithm, a comprehensive set of simulations needs are run on different video sequences with various encoding parameters. Two relevant parameters in the MPEG video standards family are the selection of GOP length and GOP

pattern. These parameters can have a huge effect on a video network. In a short GOP, a small number of frames occurs between consecutive I frames. The shortest GOP is composed of one I-frame. Thus, if a GOP is very short then the encoded video sequence will have a high number of Intra pictures. As I pictures are not as efficient at encoding data as are P pictures or B pictures, quality will be reduced and more channel bandwidth will be necessary. A short GOP is used when the video sequence contains a huge amount of motion, or when the video stream is broadcasted in a noisy environment ([437]). In a long GOP, fewer I-frames are used so fewer bandwidth is required, and more video streams can be broadcasted. GOP length and GOP pattern have a considerable impact on the visual quality of a coded video programme. However, as mentioned previously, the AVC/H.264 Standard does not define GOP. Thus, although the expression GOP in relation to H.264/AVC is widely used by the video coding community, it remains as a loosely specified expression of the key concept similar to the one provided in previous MPEG standards ([307],[437]).

In addition, the length or the pattern of a GOP is not fixed. Usually, a GOP contains one Intra coded picture (an IDR or I picture), and an arbitrary number of inter coded pictures. The selection of these parameters depends on application requirements.

Random access is vital in compressed video delivery systems such as DBV or IPTV for changing channel, bitstream splicing or trick modes ([307]). As a result, recommended procedures on how to use these parameters is central for the success of digital video applications. When a user selects a new channel, one enters the video bitstream at a random location. Pictures that use prediction based on previous broadcasted pictures cannot be correctly decoded. This problem continues until the next picture coded as Intra. Thus, encoding a picture in Intra mode at regular intervals is recommended. The distance between two Intra coded (I or IDR) pictures defines the channel-changing time. Typically, in broadcast environments, the GOP length is selected corresponds to half a second or one second. Half of a second is equal to 12 frames in Europe and 15 frames in USA (standard digital television frame rate in Europe is 25 fps, and in USA 30 fps). In DVD playback or Video on Demand services, a longer distance between I frames is used as fast channel change is not a priority.

Nevertheless, not all the pictures arriving after an Intra coded picture can be decoded properly. In some cases, pictures in one GOP use pictures in the preceding GOP as a reference. Thus, each GOP cannot be decoded fully and displayed as an independent entity. This type of GOP structure is called Open GOP. Another type of GOP structure is the Closed GOP. In this case, it can be decoded as an independent entity. A closed GOP may be obtained when only pictures that arrive after the Intra coded picture are used as reference, and if the last picture in a GOP is encoded in predicted mode similar to P type of pictures. Note that in H.264/AVC, both P and B

slices can use multiple pictures as references, so this problem is more complex than in previous standards like MPEG-2. It has been proposed, as a solution to this problem, that P slices should not be allowed to be used as reference pictures that precede the Intra coded picture ([307]). This constrain will result in a small loss of coding efficiency for those P slices that are next to the Intra coded pictures ([438]).

Another difference regarding the previous standards is that B slices can be used as reference, and thus improve the motion estimation in those pictures. Generally, B slices are used as local references (by neighbouring B pictures), so that they can be quantised to a higher degree than the P pictures ([307]). In this type of GOP pattern, more B pictures can be used. Thus, a wider motion search range is used for P pictures as they are farther apart.

As above mentioned, in some cases a low delay structure for real-time applications is required. In those cases, a GOP structure based on Intra coded slices and P coded slices is used. B slices can also be used if their references are limited from pictures in the past. A vast number of GOP structures can be formed by using different combinations of the different types of pictures. It is impossible to cover all possible combinations in this section. In our simulation the Open GOP variable was disabled. In a simple application, such as reading from a DVD, this issue is not relevant. However, in a real application environment such as digital TV, where bitstream splicing, random access, special effects (fast forward, reverse etc.) and transcoding may be performed, Close GOP is preferred.

5.3.1.3 Codec Configuration

Simulations were performed with the H.264/AVC JM version 11.0 reference software ([169],[170]). Source code was compiled with Microsoft Visual C++. Changes in the source code were performed to extract data from the input video sequence, including the computation of perceptual metrics, permitting synchronization of the different video encoders so that rate control can be performed together. Chapter 6 methods were implemented, in Matlab.

GOP Pattern	IntraPeriod	Number of B Frames	Pattern
IBBP_GOP1	10	2	IBBPBBPBBPBBPBBPBBPBBPBBPBBPBB
IBBP_GOP2	4	2	IBBPBBPBBPBB
IPPP_GOP1	4	0	IPPP
IPPP_GOP2	10	0	IPPPPPPPPP

Table 5.1 – Evaluated GOP Patterns

Table 5.1 presents the four GOP structure used in simulations. Each video test sequence was encoded with common parameters defined in Table 5.2.

Testing Platform	JM reference software encoder, version 11.0
Profile/Level IDC	(77,40) Main Profile
MV resolution	1/4 pixel
Search range	±32
Intra prediction mode	All
Inter prediction mode	All
RD-Optimization	Enabled
Hardamard	ON
Symbol Mode	CABAC
UseFME	UMHexagonS

Table 5.2 – Test Coding Conditions

The NumberReferenceFrames was set to one in the case of IPPP simulations and two in the case of IBBP simulations. Each video test sequence was encoded in two modes: Open loop (RateControlEnable Disabled, fixed QP with values ranging from 10 up to 42) and Constant Bit Rate (RateControlEnable Enabled with values of 64kbps, 128kbps, 256kbps, 384kbps, 512kbps, 640kbps, 768kbps, 1024kbps, 1536kbps, and 2048kbps). The goal was to build R-D curves (R-QP, D-QP, and R-D). During the simulations, several quality metrics were used on different occasions: Peak signal-to-noise ratio (PSNR) and the Mean Square Error (MSE), Sum of Squared Differences (SSD), Mean Absolute Difference (MAD), and Sum of Absolute Differences (SAD)

$$PSNR = 10\log_{10}\frac{255^2}{MSE} \tag{5.28}$$

$$MSE = \frac{1}{HW}SSD \tag{5.29}$$

$$SSD = \sum_{x=0}^{H-1}\sum_{y=0}^{W-1}\left(p(x,y) - \hat{p}(x,y)\right)^2 \tag{5.30}$$

$$MAD = \frac{1}{HW}SAD \tag{5.31}$$

$$SAD = \sum_{x=0}^{H-1}\sum_{y=0}^{W-1}\left|p(x,y) - \hat{p}(x,y)\right| \tag{5.32}$$

where H and W are the height and the width of the image frame, and p(x, y) and $\hat{p}(x,y)$ represent the "original" and the reconstructed image frame pixels at (x, y). These metrics have been subject to a large amount of criticism for not correlating well with HVS ([193],[439]) as they cannot signify the exact perceptual quality, since they are based on a pixel to pixel

difference calculation and ignore human perception and viewing conditions. A similar study based on SSIM and on JND is presented in the following sections.

5.3.2 *Rate-Distortion Modelling based on PSNR*

The first R-D model to be assessed was the R-D model using PSNR as the fidelity criterion. The R-D graphs obtained for the video sequences Akiyo, Foreman and Football, in an open loop, are shown in Figure 5.4 (bit rate axis is in a logarithm scale). It can be seen that a proportional relationship exists between Bit Rate and Picture Quality, and that quality depends on the video nature: for the same bit rate, low complexity sequences present higher values of quality and vice-versa. This behaviour occurs in all the different GOP patterns. Figure 5.5 provides graphic representation for R-D data at Constant Bit Rate for the same three video sequences using JM rate control. In this case, a relationship between bit rate and quality can be observed.

Annex A provides additional information regarding bit rate and picture quality as a function of the quantisation parameters (varying from 10 to 42) and temporal data (the chart is plotted for the 300 frames). First observations point to a linear relationship between the quality and quantisation parameters and a non-linear relationship between the bit rate and quantisation parameters. Nevertheless, the next step is to validate these assumptions and find the best ways to represent these relationships using mathematical models.

Frequently, data can be noisy in its nature. Thus, recognizing trends in the data is important ([440]). One of the available methods for data analysis and identifying existing trends in physical systems is curve fitting. The concept of curve fitting is rather simple: to use a function to describe a trend by minimizing the error between the selected function to fit and a set of data ([440]). The principle of least squares is applied to the fitting of a line to (x, y) data.

This principle has been used to estimate the quantiser and it is incorporated into the process of building rate-quantisation (R-Q) models using mathematical functions such as polynomial (including linear and quadratic) ([23],[25],[57],[441]), spline ([442]), logarithmic ([22],[443]), power ([24],[444]), etc. Yang et al. ([445]) proposed a more complex model that combines a logarithmic and a quadratic model. Most of the models only consider the rate function, and often implicitly assume that distortion is a linear function of the quantisation scale. This work has been extended to include D(QP) implementing several methods in order to compare their results. In fact, the goal is to model the quality versus quantisation step relationship and then to evaluate the different approaches to quality metric. It is presumed that there is an inverse relationship between quality and distortion.

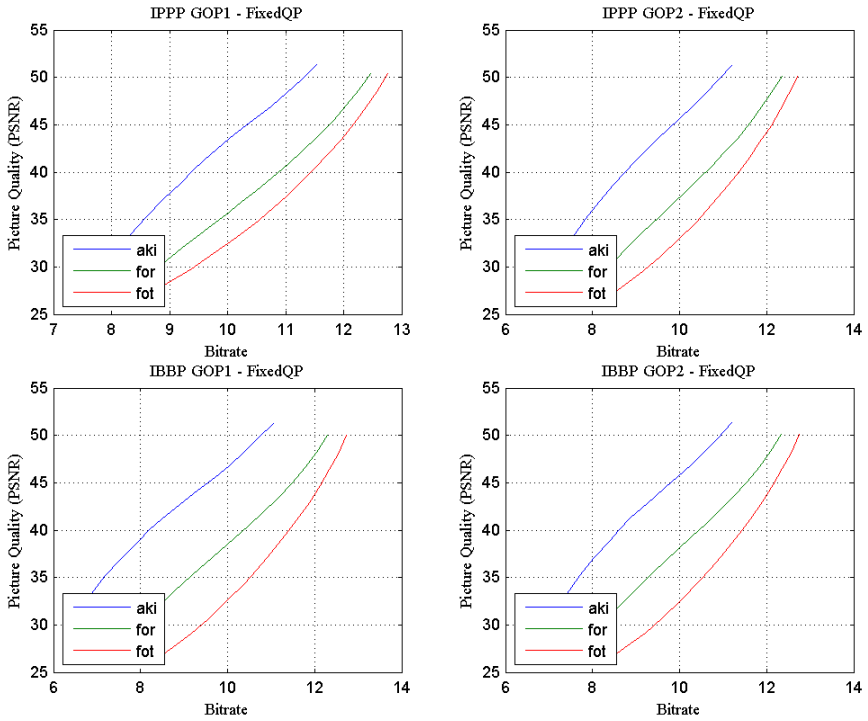


Figure 5.4 – R-PSNR curve (Akiyo, Foreman, Football; OpenLoop)

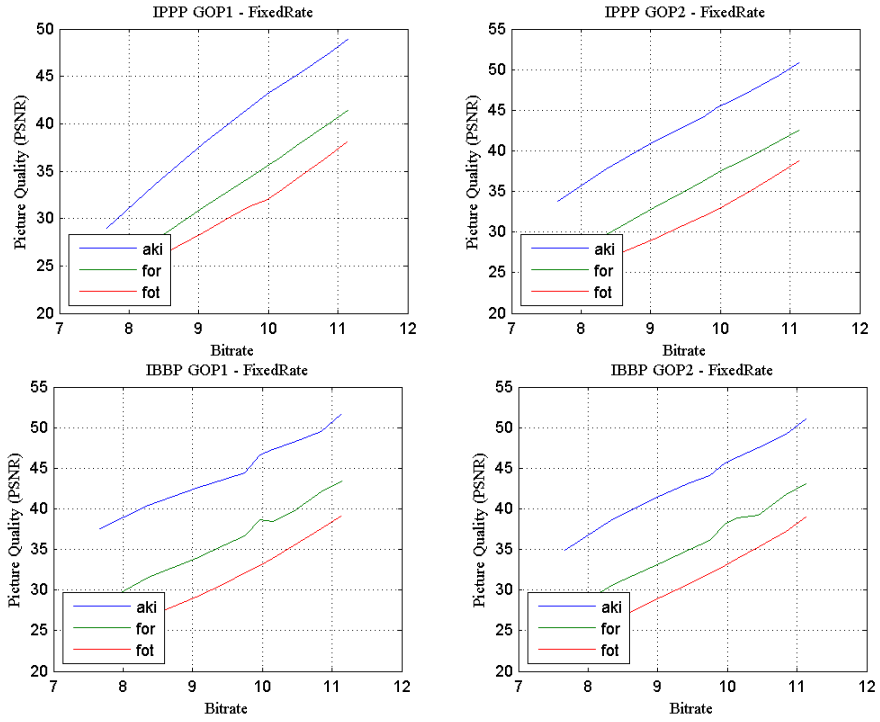


Figure 5.5 – R-PSNR curve (Akiyo, Foreman, Football; FixeRate)

Before fitting data into a function that models the relationship between two measured quantities, it is normal procedure to determine if a relationship exists between these quantities. It was decided to use the correlation method to assess the degree of probability that a relationship exists between two measured quantities ([440]). In the case of no correlation between the two quantities, then there is no tendency for the values of one quantity to increase or decrease with the values of the second quantity. To evaluate the quality of the fit, the sample correlation is used that represents the normalised measure of the strength of the linear relationship between variables ([440]):

$$r = \frac{x^T y}{\sqrt{\left(x^T x\right)\left(y^T y\right)}} \tag{5.33}$$

$$r = \frac{\sum\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\sum\left(x_i - \bar{x}\right)^2\left(y_i - \bar{y}\right)^2}} = \frac{\sum x_i y_i - \frac{\left(\sum x_i\right)\left(\sum y_i\right)}{n}}{\sqrt{\left(\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}\right)\left(\sum y_i^2 - \frac{\left(\sum y_i\right)^2}{n}\right)}} \tag{5.34}$$

where r is a matrix of correlation coefficients [440]. The sample correlation always lies in the interval from -1 to 1. A value of r near to positive one or negative one is interpreted as indicating a relatively strong relationship and r near to zero is inferred as indicating a lack of relationship. The sign of r indicates whether y tends to increase or decrease with the increase of x.

Correlation coefficients between Bits Frames and Quality Metric (PSNR)						
	IBBP GOP1			IBBP GOP2		
Sequence	I Type	P Type	B Type	I Type	P Type	B Type
Aki	0,8380	0,8470	0,9016	0,8645	0,8447	0,9034
Cgd	0,9210	0,9136	0,9595	0,9180	0,9139	0,9609
Dea	0,8853	0,8909	0,9303	0,8943	0,8878	0,9318
Flg	0,9137	0,9035	0,9349	0,9147	0,8962	0,9342
For	0,8964	0,8881	0,9197	0,8968	0,8836	0,9197
Fot	0,9588	0,9557	0,9691	0,9567	0,9550	0,9695
Hal	0,8154	0,7972	0,8589	0,8003	0,7936	0,8628
Mad	0,8797	0,8666	0,9124	0,8653	0,8645	0,9129
New	0,9554	0,9081	0,9511	0,9091	0,9128	0,9524
Par	0,9455	0,9435	0,9638	0,9461	0,9434	0,9651
Sil	0,9451	0,9412	0,9567	0,9419	0,9421	0,9576
Mcl	0,9356	0,9272	0,9470	0,9329	0,9250	0.9488

Table 5.3 – Correlation coefficients between Bits Frames and Quality Metric (PSNR) for different H.264/AVC video sequences (IBBP GOP1 and IBBP GOP2)

Correlation coefficients between Bits Frames and Quality Metric (PSNR)				
	IPPP GOP1		IPPP GOP2	
Sequence	I Type	P Type	I Type	P Type
Aki	0.8962	0.9170	0.9107	0.9065
Cgd	0.9608	0.9617	0.9615	0.9609
Dea	0.9304	0.9406	0.9354	0.9348
Flg	0.9555	0.9586	0.9567	0.9572
For	0.9268	0.9333	0.9326	0.9289
Fot	0.9659	0.9686	0.9649	0.9665
Hal	0.8540	0.8848	0.8598	0.8646
Mad	0.9019	0.9200	0.9225	0.9121
New	0.9353	0.9492	0.9526	0.9391
Par	0.9609	0.9668	0.9627	0.9630
Sil	0.9605	0.9662	0.9613	0.9621
Mcl	0.9584	0.9715	0.9615	0.9636

Table 5.4 – Correlation coefficients between Bits Frames and Quality Metric (PSNR) for different H.264/AVC video sequences (IPPP GOP1 and IPPP GOP2)

Equation (5.34) was computed for all the twelve sequences, and results were obtained according to the different Picture Type and GOP pattern (Table 5.3 and Table 5.4). Hence, the hypothesis of a relationship between PSNR and Rate was assessed. Results are very high for all the video sequences and GOP patterns, near to positive one, clearly indicating that a strong positive linear relationship is evident. The next step is thus to select which curve fitting functions should be assessed. Due to its simplicity, the first to be selected is one of the most commonly used techniques: the fitting of a straight line to a set of bivariate data generating a linear equation such as (5.35) ([440]):

$$\text{Linear} \quad y = \beta_0 + \beta_1 x \tag{5.35}$$

A natural generalization of equation (5.35) is the polynomial equation (5.36)

$$\text{Polynomial} \quad y = \beta_0 + \beta_1 x + \beta_2 x^2 + ... + \beta_k x^k \tag{5.36}$$

The goal is thus to minimise the function of k + 1 variables

$$S(\beta_0, \beta_1, \beta_2, ..., \beta_k) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left(y_i - \left(\beta_0 + \beta_1 x_i + \beta_2 x_i^{\,2} + ... + \beta_k x_i^{\,k} \right) \right)^2 \tag{5.37}$$

by selecting the coefficients $\beta_0, \beta_1, \beta_2, ..., \beta_k$ ([440]). On setting the partial derivatives of S($\beta_0, \beta_1, \beta_2, ..., \beta_k$)equal to zero and doing some simplifications, the normal equations for this least square's problem is obtained:

$$\begin{aligned} &n\beta_0 + \left(\sum x_i\right)\beta_1 + \left(\sum x_i^2\right)\beta_2 + ... + \left(\sum x_i^k\right)\beta_k = \sum y_i \\ &\left(\sum x_i\right)\beta_0 + \left(\sum x_i^2\right)\beta_1 + \left(\sum x_i^3\right)\beta_2 + ... + \left(\sum x_i^{k+1}\right)\beta_k = \sum x_i y_i \\ &\left(\sum x_i^k\right)\beta_0 + \left(\sum x_i^{k+1}\right)\beta_1 + \left(\sum x_i^{k+2}\right)\beta_2 + ... + \left(\sum x_i^{2k}\right)\beta_k = \sum x_i^k y_i \end{aligned} \quad (5.38)$$

Solving the system of k+1 linear equations presented in (5.38) it is typically possible to obtain a single set of values S($b_0, b_1, b_2, ..., b_k$) that minimises S($\beta_0, \beta_1, \beta_2, ..., \beta_k$). Polynomials are often used when a simple empirical model is required. One of the most known polynomial models is the quadratic model (also included in the study) (5.39):

$$\text{Quadratic} \quad y = \beta_0 + \beta_1 x + \beta_2 x^2 \quad (5.39)$$

It was decided, to compare with the solution available in the literature, to extend the models and thus include the logarithmic (5.40), the exponential(5.41), the power (5.42) and the linear with nonpolynomial model (LNP) (5.43).

$$\text{Logarithmic} \quad y = \beta_0 + \log x \quad (5.40)$$

$$\text{Exponential} \quad y = \beta_0 e^{\beta_1 x} \quad (5.41)$$

$$\text{Power} \quad y = \beta_0 + \beta_1 x^{\beta_2} \quad (5.42)$$

$$\text{Linear with nonpolynomial} \quad y = \beta_0 + \beta_1 e^{-x} + \beta_2 x e^{-x} \quad (5.43)$$

After selecting these six models, the average absolute error was computed when trying to model the relationship between the bit rate and quantisation parameter (QP), PSNR and quantisation parameters, and bit rate and PSNR regarding the picture type using each of the six models for all the GOP patterns. The procedure described in Figure 5.6 was implemented. Results are presented in Table 5.6 and Table 5.6. Annex B presents results for all video sequences.

Fit Method	IPPP GOP1		IPPP GOP2		IBBP GOP1			IBBP GOP2		
	I Type	P Type	I Type	P Type	I Type	P Type	B Type	I Type	P Type	B Type
Linear fit	1285	1110	2114	807	4290	1166	965	2453	1264	1051
Quadratic fit	231	154	361	128	1002	328	196	614	363	200
Exponential fit	542	505	864	358	1584	410	396	872	442	436
Logarithmic fit	996	762	1603	590	3329	980	740	1976	1065	782
Power Regression	1023	1045	1712	712	3255	747	780	1725	778	880
LNP fit	1606	2344	2998	1389	6326	802	1377	2830	856	1760

Table 5.5 – Mean Absolute Error for Rate-QP curve fitting

Fit Method	IPPP GOP1		IPPP GOP2		IBBP GOP1			IBBP GOP2		
	I Type	P Type	I Type	P Type	I Type	P Type	B Type	I Type	P Type	B Type
Linear fit	0.05	0.03	0.08	0.03	0.18	0.06	0.04	0.11	0.06	0.04
Quadratic fit	0.02	0.01	0.03	0.01	0.06	0.02	0.01	0.04	0.02	0.01
Exponential fit	0.05	0.03	0.08	0.03	0.15	0.05	0.03	0.10	0.06	0.04
Logarithmic fit	0.08	0.04	0.12	0.04	0.20	0.07	0.05	0.13	0.08	0.05
Power Regression	0.14	0.08	0.22	0.07	0.35	0.12	0.08	0.22	0.13	0.08
LNP fit	0.77	0.45	1.23	0.41	2.10	0.70	0.47	1.31	0.76	0.47

Mean Absolute Error for PSNR-QP curve fitting

Fit Method	IPPP GOP1		IPPP GOP2		IBBP GOP1			IBBP GOP2		
	I Type	P Type	I Type	P Type	I Type	P Type	BI Type	I Type	P Type	B Type
Linear fit	9789	13947	10153	11387	11411	9470	11659	10344	9421	12543
Quadratic fit	1548	1845	1576	1652	2045	1976	1914	1908	2013	1970
Exponential fit	6954	11853	7034	8854	9265	6966	9087	8261	6726	10193
Logarithmic fit	11497	17611	12097	13859	13586	10704	13852	12030	10613	15178
Power Regression	4541	7258	4312	5525	5788	4655	5934	5321	4616	6574
LNP fit	25074	50461	27787	35324	33094	19738	32635	26451	19489	38917

Mean Absolute Error for Rate-PSNR curve fitting

Table 5.6 – Mean Absolute Error for PSNR-QP and Rate-PSNR curve fitting

```
1. for each method do
2.     square error R(QP)(Picture Type) = 0;
3.     square error D(QP)(Picture Type) = 0;
4.     for each frame in the sequence do
5.         for each QP do
6.             Extract Statistics [Bits, PSNR, Picture Type];
7.         endfor
8.         Estimate the parameters of the model for R(QP) (Picture Type);
9.         Compute the square error R for each D value (Picture Type);
10.        Update the accumulative squared error R(Picture Type);
11.        Estimate the parameters of the model for D(QP) (Picture Type);
12.        Compute the square error D for each D value (Picture Type);
13.        Update the accumulative squared error D(Picture Type);
14.        Estimate the parameters of the model for R(D) (Picture Type);
15.        Compute the square error R for each D value(Picture Type);
16.        Update the accumulative squared error R_D(Picture Type);
17.    endfor
18. endfor
```

Figure 5.6 – Pseudo code for R-D model fitting

Several observations can be made from the results. Firstly, regarding accuracy, the lowest and highest results are obtained with the linear with nonpolynomial model and the quadratic approach respectively. The second observation is that the accuracy of all models varies with the level of complexity of the video source data. The accuracy of the results increases in video sequences with low complexity levels while it decreases for sequences with higher complexity values. Third observation: GOP pattern has an impact on the average of the absolute error for the different type of pictures. Nevertheless, this value, for most of the models except for the linear with nonpolynomial model, is rather small.

Sequence	Fit Method	Rate-QP		PSNR-QP		Rate - PSNR	
		I Type	P Type	I Type	P Type	I Type	P Type
Akiyo	Linear fit	237	406	0.03	0.02	2128	6295
	Quadratic fit	65	62	0.01	0.01	635	1175
	Exponential fit	79	46	0.07	0.04	761	718
	Logarithmic fit	185	269	0.11	0.07	2369	7386
	Power Regression	91	211	0.16	0.10	1024	1751
	LNP fit	329	856	0.75	0.44	5755	22485
Foreman	Linear fit	1052	1076	0.05	0.03	8368	13411
	Quadratic fit	288	209	0.01	0.01	1917	2238
	Exponential fit	320	449	0.05	0.03	2658	10807
	Logarithmic fit	866	780	0.08	0.04	9314	16117
	Power Regression	317	831	0.12	0.07	3151	7614
	LNP fit	930	1872	0.70	0.41	19274	44875
Football	Linear fit	1864	1393	0.07	0.04	13364	15256
	Quadratic fit	324	194	0.02	0.01	1931	2186
	Exponential fit	394	377	0.04	0.02	8330	15092
	Logarithmic fit	1370	981	0.05	0.03	15922	19000
	Power Regression	1191	1007	0.10	0.05	4711	9574
	LNP fit	2831	2481	0.68	0.39	38402	55186

Table 5.7 – Absolute error for Rate-QP, PSNR-QP and Rate-PSNR curve fitting (IPPP GOP1)

Sequence	Fit Method	Rate-QP		PSNR-QP		Rate - PSNR	
		I Type	P Type	I Type	P Type	I Type	P Type
Akiyo	Linear fit	462	228	0.03	0.01	2620	3879
	Quadratic fit	108	43	0.02	0.01	671	828
	Exponential fit	111	39	0.10	0.04	627	687
	Logarithmic fit	339	157	0.17	0.06	2985	4504
	Power Regression	212	112	0.25	0.09	974	1252
	LNP fit	791	451	1.18	0.40	8110	13199
Foreman	Linear fit	1776	717	0.09	0.03	8771	10188
	Quadratic fit	460	169	0.02	0.01	1804	2044
	Exponential fit	434	277	0.07	0.02	2585	6455
	Logarithmic fit	1444	550	0.11	0.04	9857	11844
	Power Regression	542	450	0.19	0.06	2680	5056
	LNP fit	1707	1045	1.11	0.37	21255	29684
Football	Linear fit	3043	1068	0.11	0.04	13687	13833
	Quadratic fit	532	179	0.03	0.01	2111	2059
	Exponential fit	664	250	0.07	0.02	8938	10504
	Logarithmic fit	2212	774	0.09	0.03	16452	16704
	Power Regression	1974	711	0.16	0.05	5008	6378
	LNP fit	4872	1722	1.09	0.36	40679	43617

Table 5.8 – Absolute error for Rate-QP, PSNR-QP and Rate-PSNR curve fitting (IPPP GOP2)

Sequence	Fit Method	Rate-QP			PSNR-QP			Rate - PSNR		
		I Type	P Type	B Type	I Type	P Type	B Type	I Type	P Type	B Type
Akiyo	Linear fit	1063	115	221	0.05	0.02	0.01	3554	1101	3211
	Quadratic fit	190	46	51	0.04	0.02	0.01	787	467	821
	Exponential fit	218	56	44	0.17	0.06	0.04	689	561	694
	Logarithmic fit	707	99	163	0.28	0.10	0.07	4164	1173	3649
	Power Regression	605	37	96	0.42	0.14	0.10	1134	683	1129
	LNP fit	2269	65	368	2.04	0.68	0.46	12646	2065	9806
Foreman	Linear fit	2736	726	767	0.14	0.04	0.03	8112	6506	9608
	Quadratic fit	894	312	209	0.06	0.02	0.01	2140	2286	2200
	Exponential fit	1123	389	286	0.11	0.05	0.03	2798	2561	5044
	Logarithmic fit	2175	634	605	0.19	0.08	0.04	9144	7028	10961
	Power Regression	1132	265	428	0.32	0.12	0.07	3557	3533	4228
	LNP fit	3396	414	970	1.92	0.66	0.43	22550	12363	25987
Football	Linear fit	5893	1564	1368	0.21	0.07	0.05	14309	12830	15554
	Quadratic fit	1038	289	241	0.07	0.03	0.02	2248	2243	2346
	Exponential fit	1779	558	369	0.12	0.04	0.03	12625	8303	11469
	Logarithmic fit	4261	1179	991	0.15	0.06	0.04	17668	15125	18796
	Power Regression	4411	1254	962	0.28	0.10	0.06	7311	4577	6842
	LNP fit	9912	2162	2207	1.95	0.67	0.43	45468	33470	47853

Table 5.9 – Absolute error for Rate-QP, PSNR-QP and Rate-PSNR curve fitting (IBBP GOP1)

Sequence	Fit Method	Rate-QP			PSNR-QP			Rate - PSNR		
		I Type	P Type	B Type	I Type	P Type	B Type	I Type	P Type	B Type
Akiyo	Linear fit	467	120	296	0.04	0.03	0.02	2452	1056	4331
	Quadratic fit	109	49	58	0.03	0.02	0.01	644	447	939
	Exponential fit	118	59	48	0.12	0.07	0.04	606	547	740
	Logarithmic fit	328	104	206	0.19	0.12	0.07	2835	1124	5009
	Power Regression	247	38	142	0.27	0.16	0.10	897	664	1358
	LNP fit	915	67	568	1.29	0.75	0.46	8218	1961	14468
Foreman	Linear fit	1559	768	861	0.08	0.04	0.03	7387	6307	10540
	Quadratic fit	565	331	206	0.04	0.02	0.01	2144	2255	2173
	Exponential fit	677	430	342	0.09	0.05	0.03	2361	2753	6879
	Logarithmic fit	1297	671	652	0.15	0.09	0.05	8159	6801	12310
	Power Regression	544	296	575	0.23	0.14	0.08	3318	3672	5214
	LNP fit	1517	431	1315	1.24	0.72	0.44	17604	11880	31947
Football	Linear fit	3228	1714	1428	0.12	0.07	0.05	13308	12846	15837
	Quadratic fit	547	328	237	0.04	0.03	0.02	2263	2280	2469
	Exponential fit	1047	609	398	0.08	0.05	0.03	10027	8032	12944
	Logarithmic fit	2368	1295	1024	0.11	0.06	0.04	16052	15114	19319
	Power Regression	2510	1357	1031	0.20	0.11	0.06	5714	4359	7924
	LNP fit	5049	2352	2398	1.27	0.73	0.44	38424	33451	51582

Table 5.10 – Absolute error for Rate-QP, PSNR-QP and Rate-PSNR curve fitting (IBBP GOP2)

The video sequence results can be analysed individually, according to model fit, picture type, and GOP pattern for the different rate-distortion-quantisation models (full data in Annex B).

Of the twelve video sequences, quadratic approach presents the best results for nine and ten video sequences when modelling respectively Intra and P pictures using IPPP GOP patterns. As for IBBP1 GOP1 and IBBP GOP2 results, quadratic approach also presents the smallest error in most cases (regarding I, P and B slice-coding, of the twelve video sequences, quadratic approach is the best results in respectively 11, 6 and 8 video sequences for IBBP GOP1 and 10, 6 and 10 for IBBP GOP2). Therefore, quadratic approach, for most cases, presents the smallest error regarding Rate-QP curve fitting.

Exponential fit and power regression present also good results, particularly for GOP patterns containing B images and for video sequences containing low to medium level of spatial and temporal complexity. In these cases, the quadratic approach is usually the second best approach. In addition, quadratic approach presents, on average, the best results when modelling Rate-PSNR. Of the twelve video sequences encoded with IPPP GOP1 pattern, quadratic performance is the best in eleven cases, for both I and P frame types. Similar results can be observed for IPPP GOP2 pattern, IBBP GOP1 and IBBP GOP2. Thus, quadratic approach is the best approach, particularly for video sequences encoded with IPPP GOP pattern.

Very good results were obtained, for all the different methods, when one analyse quality versus the quantisation parameter. Linear fit results are quite interesting, as absolute error is rather small, particularly for low complex video sequences, and the approach presents a low complexity. These results indicate that aggregate video results might be represented by the following equations:

$$R = \beta_0 + \beta_1 QP + \beta_2 QP^2 \tag{5.44}$$

$$PSNR = \beta_0' + \beta_1' \times QP + \beta_2' \times QP^2 \tag{5.45}$$

$$R = \beta_0'' + \beta_1'' \times PSNR + \beta_2'' \times PSNR^2 \tag{5.46}$$

5.3.3 *Rate-Distortion Modelling based on JND*

This section will analyse the R-QP, D-QP, and R-D using as the fidelity criterion SAD_{JND}, SSD_{JND} and PSPNR. SAD_{JND} and SSD_{JND} are determined using the following equations:

$$SAD_{JND} = \sum_{x=0}^{H-1} \sum_{y=0}^{W-1} \left| e(x,y)_{JND} \right| \tag{5.47}$$

$$SSD_{JND} = \sum_{x=0}^{H-1} \sum_{y=0}^{W-1} \left(e(x,y)_{JND} \right)^2 \tag{5.48}$$

where H and W are the height and the width of the image frame, and $e(x,y)_{JND}$ is defined by Equation (2.33) and $\delta(x,y)$ by Equation (2.34). Tests were performed using the same set of test video sequences (twelve CIF video sequences), with a duration of 10 seconds.

Simulations were performed with the JM reference software for the H.264/AVC Main profile ([317]). The JND module was implemented in Matlab. In order to be able to compare results simulations were run for all the four GOP patterns. Global simulations parameters are defined in section 5.3.1. Each video test sequence was encoded in two modes:simulation's(fixed QP with values ranging from 10 up to 42) and Constant Bit Rate (Fixed Rate - 64kbps, 128kbps, 256kbps, 384kbps, 512kbps, 640kbps, 768kbps, 1024kbps, 1536kbps, and 2048kbps).

First results are presented as R-D charts for each of three quality metrics: SAD_{JND} (Figure 5.7), SSD_{JND} (Figure 5.8) and PSPNR (Figure 5.9), for all GOP Patterns and Fixed QP.

Although simulations were performed for all the twelve sequences, results are represented for Akiyo, Football and Foreman (a sequence with low spatial and temporal complexity, a sequence

with high spatial and temporal complexity and a sequence with levels in between) in order to avoid an overflow of data.

The first observations point to a non-linear relationship between the quality and the quantisation parameter. Regarding SAD_{JND} (Figure 5.7) this relation is monotonically, strictly decreasing, regardless of the GOP pattern or video sequence. Similar behaviour is observed for SSD_{JND} (Figure 5.8). Regarding PSPNR, the relationship is monotonically, strictly increasing (Figure 5.9). It is worth noting that the relationship between PSPNR and SAD_{JND} and SSD_{JND} is inversely proportional. Therefore, PSPNR measures the quality, while SAD_{JND} and SSD_{JND} measure distortion. As observed in previous sections, Akiyo for the same quantisation parameter presents the highest quality level, the lowest distortion level and generated number of bits.

The next step is to determine if a relationship exists between these quantities. It was decided to use the correlation method to confirm the degree of probability that a relationship exists between two measured quantities ([440]) as described in more detail in Chapter 2. Average values of the correlation coefficient for rate-distortion and distortion-quantisation, regarding the three quality metrics SAD_{JND}, SSD_{JND} and PSPNR, for all the twelve video test sequences, and according to each picture type, are presented in Table 5.11, Table 5.12, Table 5.13 and Table 5.14.

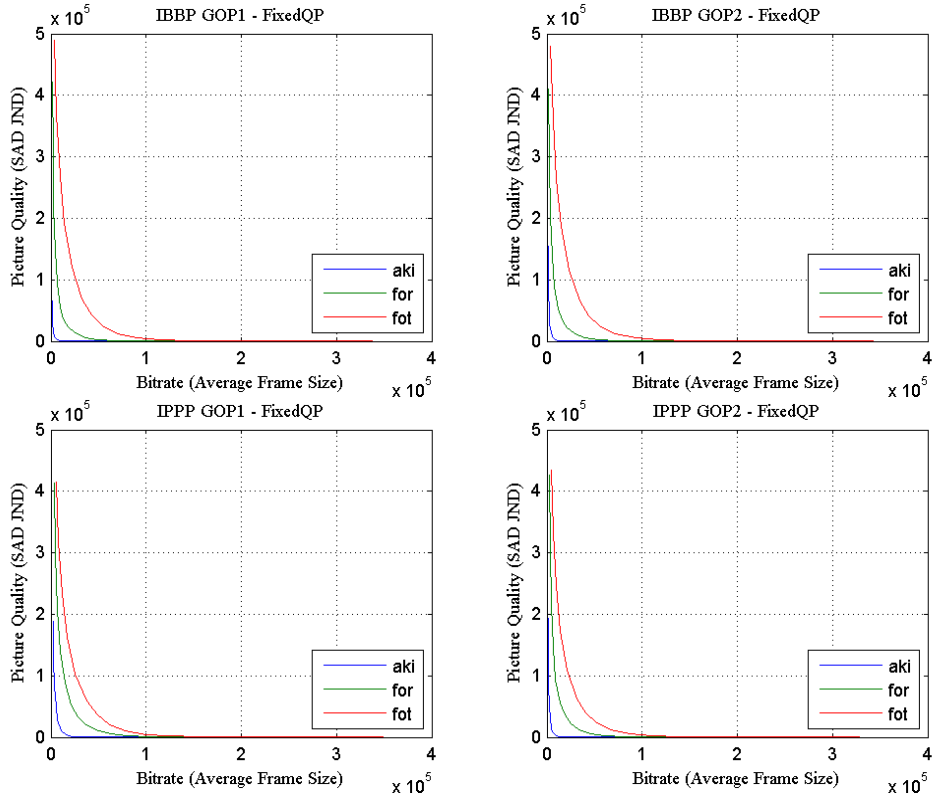


Figure 5.7 – Rate-distortion curve (SAD_JND; Akiyo, Foreman, Football)

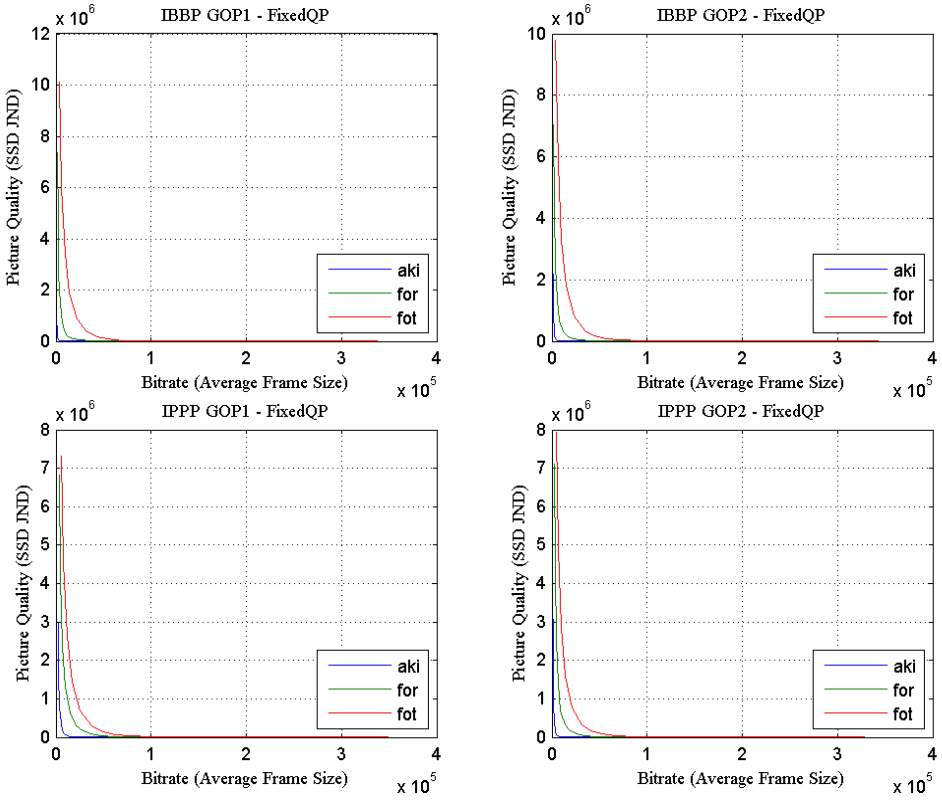


Figure 5.8 – Rate-distortion curve (SSD_JND; Akiyo, Foreman, Football)

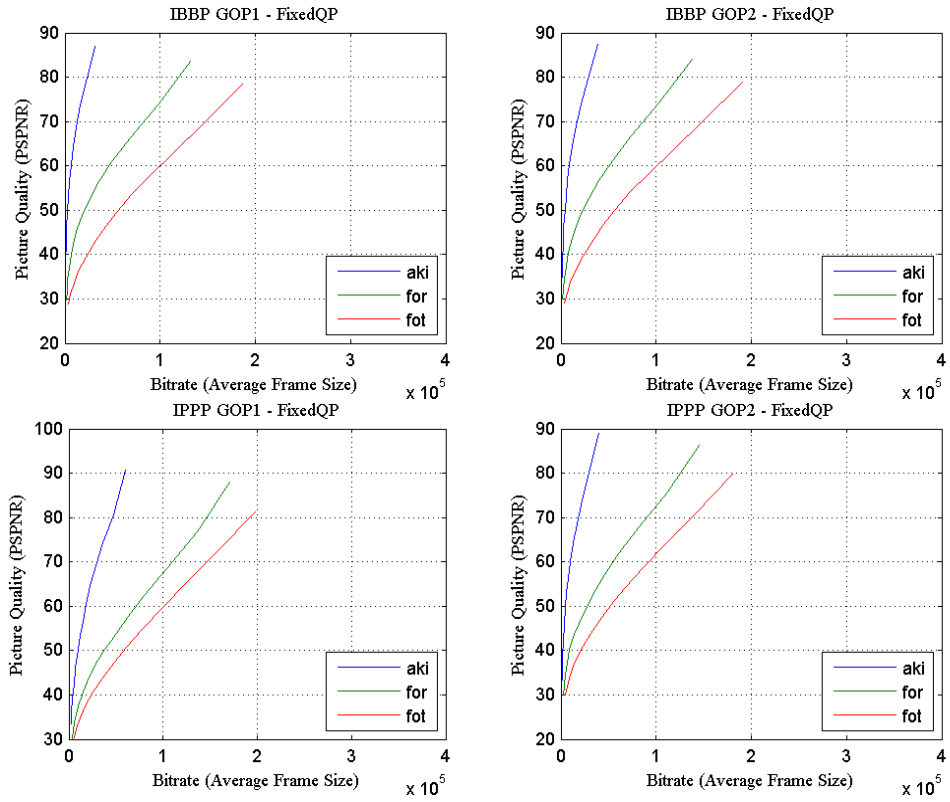


Figure 5.9 – Rate-distortion curve (PSPNR; Akiyo, Foreman, Football)

There are two ways of viewing independent variation. One is that the less distinct and related the covariation is then the greater dependence will be. Then, a value of -1.00 or 1.00 indicates full dependence, and zero correlations represent complete independence. Independence viewed in this way is called statistical independence. Two variables are then statistically independent if their correlation is zero. There is another view of independence, called linear independence that looks at dependence or independence as a matter of presence or absence, no more or less. From this point of view, two variables ranging perfectly together are linearly dependent. Hence, variables that present correlation values of -1.00 or 1.00 are linearly dependent. Otherwise, the variables are linearly independent. The first approach was followed. Table 5.11 and Table 5.12 present results regarding R-D correlation coefficients for the three quality metrics. The signal of the PSPNR results is the opposite of the SAD_{JND} and SSD_{JND}.

SAD_JND	IPPP GOP1		IPPP GOP2	
Sequence	I Type	P Type	I Type	P Type
aki	-0.7308	-0.7654	-0.7535	-0.7475
cgd	-0.8504	-0.8468	-0.8491	-0.8480
dea	-0.7662	-0.7831	-0.7776	-0.7743
flg	-0.8100	-0.8176	-0.8132	-0.8149
for	-0.7939	-0.8037	-0.8025	-0.7971
fot	-0.8665	-0.8669	-0.8647	-0.8667
hal	-0.6841	-0.7263	-0.6916	-0.6980
mad	-0.7405	-0.7712	-0.7665	-0.7555
new	-0.7740	-0.7965	-0.8014	-0.7797
par	-0.8166	-0.8252	-0.8215	-0.8203
sil	-0.8592	-0.8678	-0.8599	-0.8614
mcl	-0.8145	-0.8425	-0.8207	-0.8258

SSD_JND	IPPP GOP1		IPPP GOP2	
Sequence	I Type	P Type	I Type	P Type
aki	-0.6334	-0.6724	-0.6560	-0.6514
cgd	-0.7512	-0.7471	-0.7485	-0.7480
dea	-0.6708	-0.6902	-0.6826	-0.6797
flg	-0.7023	-0.7108	-0.7052	-0.7077
for	-0.7053	-0.7173	-0.7136	-0.7083
fot	-0.7668	-0.7684	-0.7649	-0.7677
hal	-0.5806	-0.6252	-0.5887	-0.5951
mad	-0.6483	-0.6829	-0.6730	-0.6640
new	-0.6772	-0.7028	-0.7053	-0.6833
par	-0.7165	-0.7259	-0.7222	-0.7206
sil	-0.7728	-0.7829	-0.7734	-0.7755
mcl	-0.7051	-0.7379	-0.7118	-0.7181

PSPNR	IPPP GOP1		IPPP GOP2	
Sequence	I Type	P Type	I Type	P Type
aki	0.9393	0.9530	0.9502	0.9470
cgd	0.9791	0.9786	0.9792	0.9786
dea	0.9628	0.9679	0.9663	0.9655
flg	0.9788	0.9781	0.9800	0.9792
for	0.9580	0.9615	0.9624	0.9591
fot	0.9858	0.9873	0.9854	0.9860
hal	0.8775	0.9052	0.8820	0.8870
mad	0.9367	0.9490	0.9526	0.9442
new	0.9655	0.9742	0.9783	0.9680
par	0.9851	0.9875	0.9859	0.9859
sil	0.9827	0.9863	0.9828	0.9836
mcl	0.9809	0.9877	0.9831	0.9838

Table 5.11 – Correlation coefficients R-D (SAD_JND; SSD_JND; PSPNR; IPPP GOP)

SAD_JND	IBBP GOP1			IBBP GOP2		
Sequence	I Type	P Type	B Type	I Type	P Type	B Type
aki	-0.6627	-0.6691	-0.7406	-0.6919	-0.6636	-0.7420
cgd	-0.7596	-0.7419	-0.8559	-0.7539	-0.7383	-0.8592
dea	-0.7029	-0.7051	-0.7717	-0.7109	-0.6971	-0.7727
flg	-0.7215	-0.7004	-0.7872	-0.7197	-0.6874	-0.7871
for	-0.7591	-0.7505	-0.7863	-0.7632	-0.7429	-0.7850
fot	-0.8573	-0.8478	-0.8743	-0.8495	-0.8445	-0.8742
hal	-0.6397	-0.6111	-0.6966	-0.6178	-0.6046	-0.7035
mad	-0.7135	-0.6923	-0.7605	-0.6928	-0.6891	-0.7616
new	-0.8111	-0.7441	-0.8009	-0.7443	-0.7456	-0.8015
par	-0.7846	-0.7780	-0.8249	-0.7840	-0.7743	-0.8266
sil	-0.8346	-0.8262	-0.8499	-0.8269	-0.8254	-0.8480
mcl	-0.7416	-0.7226	-0.8081	-0.7388	-0.7173	-0.8120

SSD_JND	IBBP GOP1			IBBP GOP2		
Sequence	I Type	P Type	B Type	I Type	P Type	B Type
aki	-0.5680	-0.5733	-0.6465	-0.5965	-0.5670	-0.6481
cgd	-0.6492	-0.6291	-0.7615	-0.6423	-0.6253	-0.7667
dea	-0.6043	-0.6061	-0.6784	-0.6125	-0.5988	-0.6801
flg	-0.6078	-0.5852	-0.6793	-0.6055	-0.5735	-0.6813
for	-0.6706	-0.6609	-0.6960	-0.6750	-0.6529	-0.6944
fot	-0.7543	-0.7450	-0.7679	-0.7467	-0.7412	-0.7677
hal	-0.5414	-0.5122	-0.6012	-0.5188	-0.5049	-0.6082
mad	-0.6187	-0.5991	-0.6710	-0.6013	-0.5966	-0.6741
new	-0.7162	-0.6475	-0.7054	-0.6493	-0.6503	-0.7073
par	-0.6801	-0.6731	-0.7266	-0.6792	-0.6686	-0.7286
sil	-0.7429	-0.7326	-0.7603	-0.7338	-0.7324	-0.7578
mcl	-0.6307	-0.6082	-0.7037	-0.6257	-0.6015	-0.7092

PSPNR	IBBP GOP1			IBBP GOP2		
Sequence	I Type	P Type	B Type	I Type	P Type	B Type
aki	0.8873	0.8975	0.9435	0.9118	0.8990	0.9446
cgd	0.9366	0.9294	0.9795	0.9332	0.9312	0.9800
dea	0.9218	0.9296	0.9632	0.9313	0.9281	0.9636
flg	0.9352	0.9272	0.9638	0.9349	0.9234	0.9616
for	0.9341	0.9308	0.9488	0.9376	0.9294	0.9480
fot	0.9811	0.9814	0.9858	0.9814	0.9816	0.9861
hal	0.8352	0.8148	0.8828	0.8176	0.8114	0.8863
mad	0.9161	0.9070	0.9440	0.9030	0.9075	0.9439
new	0.9808	0.9416	0.9775	0.9438	0.9474	0.9780
par	0.9717	0.9713	0.9877	0.9727	0.9720	0.9884
sil	0.9726	0.9701	0.9800	0.9713	0.9723	0.9814
mcl	0.9444	0.9391	0.9758	0.9429	0.9386	0.9753

Table 5.12 – Correlation coefficients R-D (SAD_JND; SSD_JND; PSPNR; IBBP GOP)

SAD_JND	IPPP GOP1		IPPP GOP2	
Sequence	I Type	P Type	I Type	P Type
aki	0.9440	0.9424	0.9471	0.9442
cgd	0.9782	0.9733	0.9767	0.9760
dea	0.9460	0.9429	0.9482	0.9460
flg	0.9595	0.9527	0.9576	0.9574
for	0.9640	0.9621	0.9659	0.9639
fot	0.9743	0.9726	0.9728	0.9742
hal	0.9416	0.9393	0.9421	0.9409
mad	0.9525	0.9515	0.9546	0.9530
new	0.9372	0.9349	0.9400	0.9371
par	0.9460	0.9429	0.9468	0.9453
sil	0.9792	0.9789	0.9793	0.9791
mcl	0.9608	0.9561	0.9594	0.9594

SSD_JND	IPPP GOP1		IPPP GOP2	
Sequence	I Type	P Type	I Type	P Type
aki	0.8803	0.8778	0.8837	0.8802
cgd	0.9267	0.9175	0.9231	0.9225
dea	0.8842	0.8799	0.8868	0.8842
flg	0.8939	0.8818	0.8901	0.8900
for	0.9139	0.9111	0.9165	0.9132
fot	0.9167	0.9142	0.9142	0.9169
hal	0.8725	0.8692	0.8732	0.8714
mad	0.8946	0.8931	0.8967	0.8951
new	0.8684	0.8658	0.8720	0.8685
par	0.8782	0.8729	0.8793	0.8769
sil	0.9340	0.9335	0.9341	0.9338
mcl	0.8949	0.8874	0.8925	0.8928

PSPNR	IPPP GOP1		IPPP GOP2	
Sequence	I Type	P Type	I Type	P Type
aki	-0.9950	-0.9950	-0.9943	-0.9947
cgd	-0.9896	-0.9897	-0.9900	-0.9899
dea	-0.9935	-0.9933	-0.9931	-0.9932
flg	-0.9898	-0.9905	-0.9904	-0.9902
for	-0.9925	-0.9918	-0.9917	-0.9922
fot	-0.9897	-0.9878	-0.9898	-0.9889
hal	-0.9996	-0.9996	-0.9995	-0.9996
mad	-0.9945	-0.9941	-0.9936	-0.9940
new	-0.9954	-0.9952	-0.9947	-0.9951
par	-0.9927	-0.9924	-0.9928	-0.9927
sil	-0.9875	-0.9864	-0.9877	-0.9872
mcl	-0.9893	-0.9899	-0.9897	-0.9895

Table 5.13 – Correlation coefficients D-Q (SAD_JND; SSD_JND; PSPNR; IPPP GOP)

SAD_JND	IBBP GOP1			IBBP GOP2		
Sequence	I Type	P Type	B Type	I Type	P Type	B Type
aki	0.9460	0.9449	0.9463	0.9437	0.9414	0.9433
cgd	0.9740	0.9735	0.9835	0.9750	0.9716	0.9838
dea	0.9482	0.9463	0.9489	0.9462	0.9433	0.9468
flg	0.9531	0.9556	0.9625	0.9580	0.9534	0.9623
for	0.9683	0.9681	0.9647	0.9677	0.9656	0.9626
fot	0.9769	0.9732	0.9756	0.9729	0.9726	0.9753
hal	0.9394	0.9397	0.9430	0.9399	0.9380	0.9427
mad	0.9545	0.9530	0.9566	0.9527	0.9497	0.9539
new	0.9408	0.9383	0.9399	0.9384	0.9355	0.9381
par	0.9464	0.9454	0.9479	0.9457	0.9431	0.9469
sil	0.9786	0.9786	0.9778	0.9783	0.9776	0.9762
mcl	0.9584	0.9590	0.9717	0.9628	0.9569	0.9714

SSD_JND	IBBP GOP1			IBBP GOP2		
Sequence	I Type	P Type	B Type	I Type	P Type	B Type
aki	0.8821	0.8812	0.8845	0.8794	0.8763	0.8803
cgd	0.9179	0.9171	0.9382	0.9198	0.9140	0.9395
dea	0.8854	0.8836	0.8890	0.8834	0.8796	0.8863
flg	0.8828	0.8869	0.8975	0.8912	0.8841	0.8980
for	0.9211	0.9200	0.9129	0.9201	0.9159	0.9092
fot	0.9197	0.9148	0.9129	0.9139	0.9135	0.9121
hal	0.8706	0.8710	0.8792	0.8710	0.8678	0.8784
mad	0.8954	0.8957	0.9018	0.8959	0.8914	0.8991
new	0.8729	0.8699	0.8723	0.8709	0.8669	0.8705
par	0.8774	0.8766	0.8810	0.8767	0.8728	0.8796
sil	0.9316	0.9316	0.9303	0.9315	0.9304	0.9275
mcl	0.8931	0.8934	0.9148	0.8993	0.8889	0.9146

PSPNR	IBBP GOP1			IBBP GOP2		
Sequence	I Type	P Type	B Type	I Type	P Type	B Type
aki	-0.9941	-0.9942	-0.9939	-0.9947	-0.9945	-0.9946
cgd	-0.9899	-0.9902	-0.9826	-0.9900	-0.9899	-0.9822
dea	-0.9931	-0.9934	-0.9926	-0.9936	-0.9933	-0.9930
flg	-0.9917	-0.9911	-0.9871	-0.9907	-0.9905	-0.9869
for	-0.9908	-0.9908	-0.9930	-0.9915	-0.9906	-0.9937
fot	-0.9859	-0.9871	-0.9899	-0.9878	-0.9858	-0.9900
hal	-0.9996	-0.9996	-0.9994	-0.9997	-0.9997	-0.9995
mad	-0.9933	-0.9938	-0.9937	-0.9942	-0.9943	-0.9946
new	-0.9945	-0.9948	-0.9948	-0.9947	-0.9948	-0.9950
par	-0.9926	-0.9927	-0.9913	-0.9929	-0.9925	-0.9913
sil	-0.9872	-0.9870	-0.9872	-0.9867	-0.9858	-0.9862
mcl	-0.9903	-0.9901	-0.9795	-0.9894	-0.9900	-0.9795

Table 5.14 – Correlation coefficients D-Q (SAD_JND; SSD_JND; PSPNR; IBBP GOP)

The highest results are obtained for PSPNR (typically 0.96 IPPP and 0.94 IBBP) followed by SAD_{JND} (typically -0.8 IPPP and -0.75 IBBP) and finally SSD_{JND} (typically -0.7 IPPP and -0.65 IBBP). PSPNR coefficients' values with the exception of the news sequence are above 0.9. Thus, the corresponding variables, Rate and PSPNR, closely vary together in the same direction, the higher the rate, the higher the quality.

GOP Patterns has an influence on R-D results. A reduction in average correlation, roughly from 0.05 up to 0.1, in absolutes terms, can be observed in frames of the type Intra and Predicted in IBBP GOP video programmes when compared with IPPP GOP video programmes. This fact is observed in all quality metrics.

The D-Q correlation coefficients results, per video test sequence, and for all the different type of GOP Patterns, are presented in Table 5.13 and Table 5.14. The correlation values are very high for all the metrics, especially for PSPNR (values ranging between -0.98 and -0.99). SSD_{JND} presents the lowest values, although in absolute the values are still high (ranging between 0.86 and 0.94). In all the cases, the effect of GOP pattern is rather small.

The correlation regarding PSPNR is negative (the higher the quantisation, the lower the quality) and for SSD_{JND} and SAD_{JND} it is positive (the higher the quantisation the higher the distortion). The direction results are as expected due to the nature of the quantisation process (the higher the value of the quantisation step size, the coarser the signal is encoded and fewer bits are necessary to encode data). PSPNR results are very promising for both R-D and D-Q).

The next step is to evaluate how the different curve fitting approaches model the rate-distortion relationship (SAD_{JND}, SSD_{JND} and PSPNR). Simulations followed the proceedings defined in previous section and thus six methods were selected and assessed: linear (Equation (5.35)), quadratic (Equation (5.39)), exponential (Equation (5.41)), logarithmic (Equation (5.40)), power (Equation (5.42)) and linear with nonpolynomial model (Equation (5.43)). Tests were performed for all 12 sequences and four GOP patterns, and the results are presented according to picture type.

SAD_JND	IPPP GOP1		IPPP GOP2		IBBP GOP1			IBBP GOP2		
	I Type	P Type	I Type	P Type	I Type	P Type	BI Type	I Type	P Type	B Type
Linear fit	19365	36488	21165	26128	25005	16576	23741	20700	16444	27528
Quadratic fit	11300	20539	12295	14921	15103	10879	13241	12794	10948	15003
Exponential fit	13667	26086	14950	18538	19189	13636	16291	16210	13733	18858
Logarithmic fit	5843	6810	5812	6152	7631	7223	6786	7366	7089	7240
Power Regression	14537	25721	15411	18970	18082	12403	19993	15398	12111	22155
LNP fit	34970	73764	39254	50626	46102	25499	45822	35929	25087	55396
SSD_JND	IPPP GOP1		IPPP GOP2		IBBP GOP1			IBBP GOP2		
	I Type	P Type	I Type	P Type	I Type	P Type	BI Type	I Type	P Type	B Type
Linear fit	23805	46542	26257	32846	30783	19337	29735	25010	19123	34875
Quadratic fit	17066	32176	18719	23036	22324	15045	20562	18542	14971	23665
Exponential fit	21159	40639	23223	28847	28198	18900	25943	23440	18772	30151
Logarithmic fit	6797	8012	6813	7239	8493	7989	7716	8192	7865	8203
Power Regression	12576	22268	13337	16405	15487	10362	17884	13032	10110	19862
LNP fit	34970	73764	39254	50626	46102	25499	45822	35929	25087	55396
PSPNR	IPPP GOP1		IPPP GOP2		IBBP GOP1			IBBP GOP2		
	I Type	P Type	I Type	P Type	I Type	P Type	BI Type	I Type	P Type	B Type
Linear fit	6797	8012	6813	7239	8493	7989	7716	8192	7865	8203
Quadratic fit	1357	2327	1424	1709	2062	1562	2233	1722	1585	2555
Exponential fit	12576	22268	13337	16405	15487	10362	17884	13032	10110	19862
Logarithmic fit	9424	13123	9740	10838	11478	9882	10830	10606	9737	11700
Power Regression	7744	13685	8003	10064	9016	5861	11626	7488	5788	12960
LNP fit	29377	61170	32663	42138	38845	22655	37941	31009	22690	45692

Table 5.15 – Average Absolute Error R-D (PSPNR; Picture Type; GOP Pattern)

Table 5.15 shows the average absolute error, for the complete set of video test sequences, when modelling R-D data. Regarding SAD_{JND} the best method is logarithmic followed by quadratic, power and exponential methods with similar results. GOP pattern has an impact on picture type error, but the impact depends on the method and GOP pattern. As for SSD_{JND} the best method is also the logarithmic followed by power regression (a clear second best). Results differ regarding PSPNR. With PSPNR, the best method is the quadratic method. The second best for PSPNR is the linear approach, but the difference to the quadratic method is huge. The difference is larger than that observed for the second best methods of SAD_{JND} and SSD_{JND}. PSPNR regarding R-D presents a similar behaviour that has been observed in Chapter 2 with PSNR. As for SAD_{JND} and SSD_{JND}, perceptual versions of the SAD and SSD metrics, their performance is within that found in the literature concerning SAD and SSD metrics ([446]).

SAD_JND	IPPP GOP1		IPPP GOP2		IBBP GOP1			IBBP GOP2		
	I Type	P Type	I Type	P Type	I Type	P Type	B Type	I Type	P Type	B Type
Linear fit	9767	5738	16057	5256	26576	8770	5851	16120	9553	5744
Quadratic fit	1176	725	1866	638	3038	1008	774	1847	1183	820
Exponential fit	12028	7175	20256	6556	33877	11020	7897	20179	12234	7506
Logarithmic fit	12341	7180	20295	6626	33478	11042	7474	20338	11949	7309
Power Regression	6506	3910	11018	3561	18589	6010	4534	11037	6747	4282
LNP fit	27993	15897	46074	14944	75377	24832	17443	45955	26440	16909
SSD_JND	IPPP GOP1		IPPP GOP2		IBBP GOP1			IBBP GOP2		
	I Type	P Type	I Type	P Type	I Type	P Type	B Type	I Type	P Type	B Type
Linear fit	268941	151361	455867	144622	724372	235126	172493	429653	245274	164271
Quadratic fit	70150	41828	118293	38139	190467	61942	44504	111742	67165	43527
Exponential fit	273677	154472	473965	148610	749208	240234	195279	438715	254548	180204
Logarithmic fit	308781	172922	523158	165837	829917	269445	199103	492935	280230	189370
Power Regression	120095	67185	208182	65182	327147	104535	96241	192779	111944	88390
LNP fit	531149	292663	898379	284079	1417192	460506	349004	845644	474328	330705
PSPNR	IPPP GOP1		IPPP GOP2		IBBP GOP1			IBBP GOP2		
	I Type	P Type	I Type	P Type	I Type	P Type	B Type	I Type	P Type	B Type
Linear fit	0.30	0.18	0.48	0.16	0.86	0.29	0.20	0.53	0.32	0.20
Quadratic fit	0.05	0.03	0.08	0.03	0.15	0.05	0.04	0.09	0.06	0.05
Exponential fit	0.13	0.08	0.21	0.07	0.39	0.13	0.10	0.24	0.15	0.10
Logarithmic fit	0.13	0.08	0.22	0.07	0.40	0.13	0.10	0.24	0.15	0.10
Power Regression	0.12	0.08	0.19	0.07	0.35	0.11	0.08	0.21	0.13	0.08
LNP fit	1.36	0.81	2.17	0.73	3.79	1.27	0.82	2.38	1.42	0.82

Table 5.16 – Average Absolute Error D-QP (Picture Type; GOP Pattern)

Table 5.16 contains the aggregate results for the full video test sequence of the average absolute error, per frame type, for the six models. The best method to model D-QP in the case of SAD_{JND}, SSD_{JND} and PSPNR, is the quadratic method.

Results depend on GOP pattern: in an IBBP GOP video sequence, the error is higher in the Intra frames, and lower in the Interpolated frames. Interpolated frames present higher quality for the same quantisation step due to the use of motion estimation techniques.

Results from R-Q and D-QP show that PSPNR can be modelled by a quadratic function while for SAD_{JND} and SSD_{JND} quadratic is the best approach for D-QP and logarithmic for R-D. Additional information regarding the results of the individual video sequences Akiyo, Foreman and Football are presented in Annex C. These sequences provide an excellent example of how the results vary within the twelve video test sequences.

5.3.4 *Rate-Distortion Modeling based on SSIM*

In this section, it will be characterised the rate-distortion (R-D) and distortion-quantisation (*D-Q*) functions using SSIM picture quality metric. All simulations have been performed using JM H.264/AVC reference software ([169],[170]). Teste sequences and encoder configurations are defined in Section 5.3.1. Before starting the RD modeling, it will be provided some information regarding the use of SSIM metric within H.264/AVC coding for open loop and CBR coding. An SSIM value is between 0 and 1, with 1 meaning perfect quality. SSIM results have been scaled to make it more readable using the following equation:

$$100 \times \mathrm{SSIM}^{8} \tag{5.49}$$

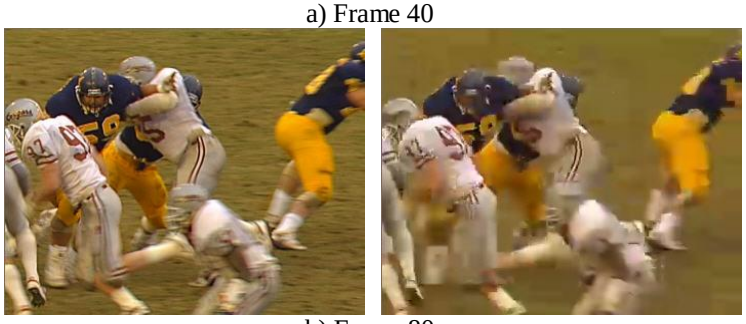
Table 5.17 presents the average SSIM for three of the sequences (Akiyo, Foreman and Football) encoded Open Loop (fixed QP) regarding the four GOP Patterns.

Akiyo	QP12	QP18	QP24	QP30	QP36	QP42
IBBP_GOP1	95.18	90.54	85.29	76.20	61.47	42.28
IBBP_GOP2	95.20	90.53	85.34	76.26	61.61	42.32
IPPP_GOP1	95.26	90.48	85.21	75.93	60.93	41.99
IPPP_GOP2	95.17	90.41	85.10	75.76	60.75	41.77
Foreman						
IBBP_GOP1	94.90	87.61	75.28	57.39	36.99	20.08
IBBP_GOP2	94.99	87.85	76.05	58.41	38.29	20.74
IPPP_GOP1	95.33	88.38	75.92	58.18	38.43	21.17
IPPP_GOP2	95.10	88.03	75.34	57.45	37.55	20.48
Football						
IBBP_GOP1	94.95	86.34	70.37	44.76	20.57	9.55
IBBP_GOP2	95.02	86.52	71.15	46.07	21.43	9.88
IPPP_GOP1	95.32	87.43	72.53	48.51	24.68	11.78
IPPP_GOP2	95.06	86.91	71.30	47.02	23.39	11.17

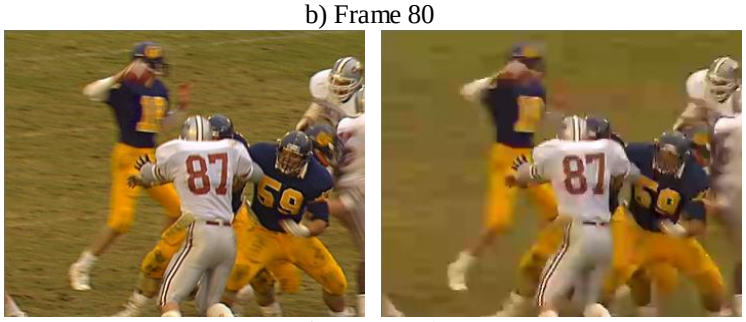
Table 5.17 – Average SSIM in Open Loop for different GOP Patterns

Results are relatively independent of the GOP Pattern. One can observe that quantisation parameters have a major impact on the final quality of the encoded video test sequence. Although it can be observe that quality variations with the quantisation depend very much on the content itself, it is possible to infer the variation of perceived quality. For example, consider Football sequence coding distortions at fixed quantisation (QP42). In terms of subjective quality, distortion is perceptible during all the sequence. Nevertheless, average SNRY is 27.5 dB with higher values in the beginning of the sequence (Figure 5.10). SSIM offers a better representation of how a typical viewer would assess this sequence.

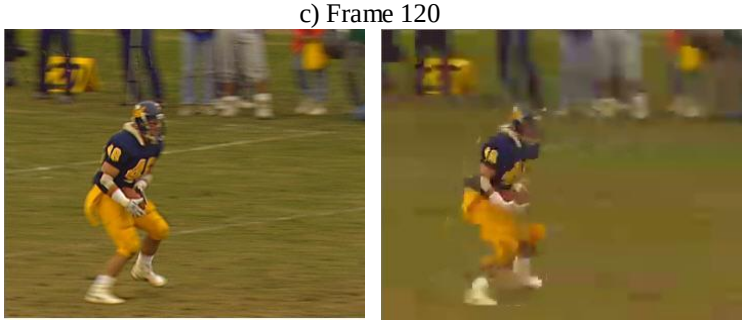
a) Frame 40



b) Frame 80



c) Frame 120



d) Frame 160

e) Frame 200

Figure 5.10 – Original and Reconstructed Frames – Open Loop (QP42 - IPPP GOP1)

Akiyo	256kbps	512kbps	768kbps	1024kbps	1536kbps	2048kbps
IBBP_GOP1	85.04	87.86	92.92	93.78	94.84	96.56
IBBP_GOP2	83.89	87.76	92.01	93.32	94.71	96.34
IPPP_GOP1	72.45	83.75	88.08	90.08	92.65	94.49
IPPP_GOP2	82.67	87.84	91.30	92.87	94.74	96.17
Foreman						
IBBP_GOP1	50.40	65.18	74.08	78.09	85.39	88.07
IBBP_GOP2	47.49	62.92	75.60	76.78	84.48	87.44
IPPP_GOP1	34.11	51.64	62.20	69.15	77.53	82.41
IPPP_GOP2	45.22	62.03	71.47	76.62	82.34	85.91
Football						
IBBP_GOP1	17.53	31.75	45.14	53.85	65.46	72.69
IBBP_GOP2	16.68	31.39	45.06	53.98	64.19	72.57
IPPP_GOP1	14.02	25.47	35.21	44.40	58.79	67.64
IPPP_GOP2	16.36	29.64	42.14	51.02	63.73	71.93

Table 5.18 – Average SSIM in CBR mode for different GOP Patterns

Table 5.18 illustrates the average SSIM value, for Akiyo, Foreman and Football video sequences, encoded at Constant Bit Rate using JM software, with four different GOP patterns. The impact of the GOP pattern in terms of quality is more visible than in PSNR; IBBP GOP patterns present higher levels of quality. It is also possible to infer the coding complexity from results (football is the most complex spatial and temporal sequence obtaining and the hardest to encode).

Regarding R-D modeling, each video test sequence was encoded in two modes: Open loop (fixed QP with values ranging from 10 up to 42) and Constant Bit Rate (Fixed Rate - 64kbps, 128kbps, 256kbps, 384kbps, 512kbps, 640kbps, 768kbps, 1024kbps, 1536kbps, and 2048kbps).

Figure 5.11 plots R-D curves, using SSIM, for all GOP Patterns and Fixed QP. It was selected three video sequences, Akiyo, Football and Foreman (a sequence with low spatial and temporal complexity, a sequence with high spatial and temporal complexity and a sequence with levels in between), to illustrate typical R-D charts, although simulations for the twelve video test sequences have been performed.

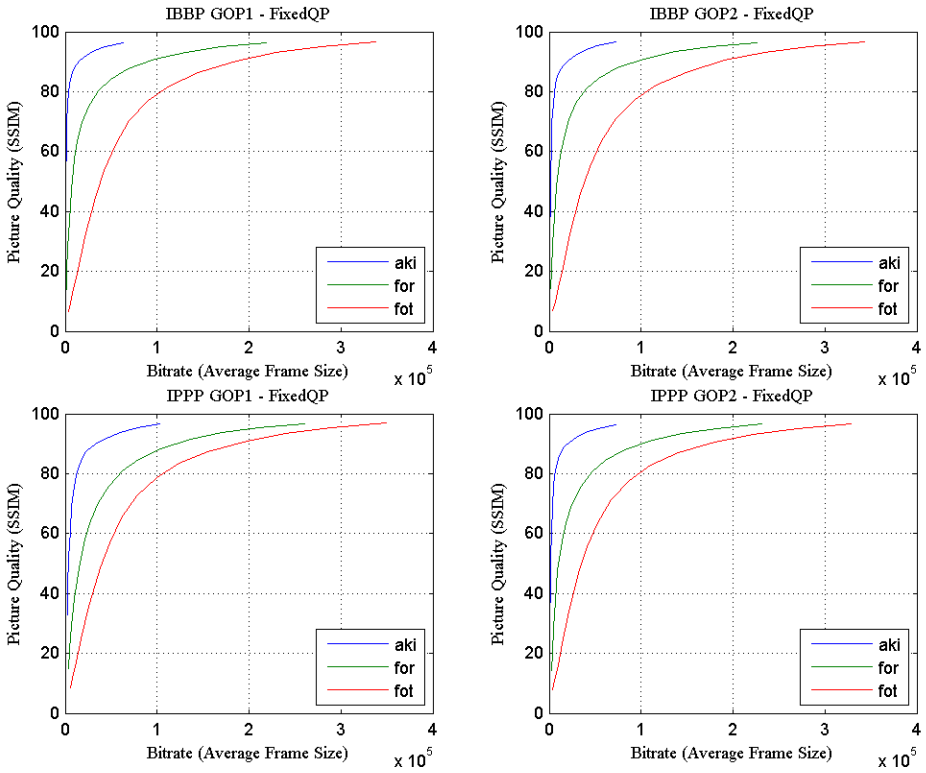


Figure 5.11 – Rate-Distortion Curve (SSIM; Akiyo, Foreman; Football)

As it can be seen from the Figure 5.11, the shape of the R-D curve starts by a sharp increase of the quality until it reaches a high value. After this point, the ratio between the quality variation and bit rate variation decreases significantly. This behaviour can be observed regardless of the GOP Pattern. This fact is less observable in higher spatial and temporal complexity, such as football. At the same time, this behaviour can be exploited in joint video coding scenarios. As an example, one can analyse the bit rate for one of the GOP Patterns charts when the level of SSIM quality equals 80. One can observe that the required number of bits to obtain this quality level varies extensively depending on the content nature. In a joint video coding scenario, this means that redistributing bandwidth among video sources can be use to maximize the minimum level of quality of the broadcast programmes.

	IPPP GOP1		IPPP GOP2	
Sequence	I Type	P Type	I Type	P Type
aki	0.8382	0.8662	0.8538	0.8509
cgd	0.9591	0.9546	0.9619	0.9583
dea	0.8812	0.8945	0.8888	0.8874
flg	0.9058	0.9119	0.9133	0.9111
for	0.9170	0.9226	0.9233	0.9186
fot	0.9756	0.9757	0.9755	0.9759
hal	0.8070	0.8452	0.8113	0.8191
mad	0.8527	0.8762	0.8751	0.8646
new	0.8922	0.9101	0.9129	0.8968
par	0.9397	0.9462	0.9426	0.9421
sil	0.9545	0.9595	0.9558	0.9560
mcl	0.9120	0.9333	0.9199	0.9214

Table 5.19 – Correlation coefficients R-D (SSIM; IPPP GOPs)

	IBBP GOP1			IBBP GOP2		
Sequence	I Type	P Type	B Type	I Type	P Type	B Type
aki	0.7670	0.7767	0.8398	0.7975	0.7764	0.8435
cgd	0.9017	0.8895	0.9618	0.8975	0.8854	0.9612
dea	0.8237	0.8290	0.8824	0.8330	0.8242	0.8838
flg	0.8408	0.8214	0.8849	0.8330	0.8077	0.8797
for	0.8867	0.8761	0.9075	0.8849	0.8715	0.9060
fot	0.9738	0.9688	0.9807	0.9687	0.9657	0.9799
hal	0.7649	0.7425	0.8012	0.7466	0.7428	0.8075
mad	0.8242	0.8078	0.8635	0.8065	0.8085	0.8656
new	0.9164	0.8621	0.9111	0.8622	0.8668	0.9131
par	0.9178	0.9143	0.9426	0.9178	0.9137	0.9439
sil	0.9335	0.9287	0.9466	0.9288	0.9283	0.9457
mcl	0.8520	0.8357	0.8960	0.8437	0.8333	0.8949

Table 5.20 – Correlation coefficients R-D (SSIM; IBBP GOPs)

Table 5.19 and Table 5.20 assess the correlation between Rate and SSIM. All the sequences from the video test set were analysed, and simulations for all the GOP patterns were performed. Results show a strong relation between Rate and SSIM metric. Concerning the results of video sequences with IPPP GOP patterns, they vary between 0.80 and 0.98 and in the case of IBBP GOP pattern results vary between 0.74 and 0.98. These results are only exceeded by PSPNR (0.96 for IPPP GOPs and between 0.93 and 0.96 for IBBP GOPs). Intra and Predicted frames decrease their results in sequences encoded with a GOP pattern IBBP. This observation is in line with the results of using JND based quality metrics.

	IPPP GOP1		IPPP GOP2	
Sequence	I Type	P Type	I Type	P Type
Aki	-0.9901	-0.9900	-0.9905	-0.9901
Cgd	-0.9946	-0.9967	-0.9935	-0.9948
Dea	-0.9938	-0.9934	-0.9943	-0.9939
Flg	-0.9951	-0.9940	-0.9954	-0.9950
For	-0.9985	-0.9986	-0.9981	-0.9986
Fot	-0.9900	-0.9916	-0.9892	-0.9895
Hal	-0.9901	-0.9909	-0.9900	-0.9904
Mad	-0.9949	-0.9949	-0.9956	-0.9952
New	-0.9928	-0.9926	-0.9934	-0.9928
Par	-0.9981	-0.9980	-0.9982	-0.9981
Sil	-0.9978	-0.9979	-0.9975	-0.9978
Mcl	-0.9961	-0.9955	-0.9964	-0.9961

Table 5.21 – Correlation coefficients D-QP (SSIM; IPPP GOP1, IPPP GOP2)

	IBBP GOP1			IBBP GOP2		
Sequence	I Type	P Type	B Type	I Type	P Type	B Type
Aki	-0.9896	-0.9895	-0.9895	-0.9892	-0.9891	-0.9889
Cgd	-0.9948	-0.9953	-0.9912	-0.9950	-0.9963	-0.9919
Dea	-0.9940	-0.9936	-0.9943	-0.9935	-0.9931	-0.9939
Flg	-0.9945	-0.9947	-0.9962	-0.9948	-0.9938	-0.9956
For	-0.9983	-0.9983	-0.9982	-0.9985	-0.9986	-0.9985
Fot	-0.9862	-0.9886	-0.9849	-0.9892	-0.9903	-0.9865
Hal	-0.9897	-0.9900	-0.9867	-0.9900	-0.9907	-0.9868
Mad	-0.9948	-0.9946	-0.9948	-0.9944	-0.9941	-0.9943
New	-0.9926	-0.9925	-0.9929	-0.9926	-0.9923	-0.9928
Par	-0.9982	-0.9981	-0.9982	-0.9981	-0.9979	-0.9980
Sil	-0.9984	-0.9985	-0.9981	-0.9986	-0.9987	-0.9984
Mcl	-0.9956	-0.9952	-0.9978	-0.9956	-0.9949	-0.9972

Table 5.22 – Correlation coefficients D-QP (SSIM; IBBP GOP1, IBBP GOP2)

Correlation coefficients between SSIM and quantisation are very high; near negative one (average value for all the video sequences is about -0.994). This fact points to a strong relation between these two variables. Negative correlation results due to the nature of the quantisation process. Higher quantisation parameters imply larger coding errors that result in smaller values of quality. This is the case where it can be observed the strongest correlation detected through the study. In addition, results are independent of the spatial and temporal complexity of the video sequences. The impact of picture type or GOP pattern is very small, and results are very homogeneous. After these results, next step is the evaluation of the best function that allows modelling R-D and D-QP.

Fit Method	IPPP GOP1		IPPP GOP2		IBBP GOP1			IBBP GOP2		
	I Type	P Type	I Type	P Type	I Type	P Type	BI Type	I Type	P Type	B Type
Linear fit	12229	20024	12817	15193	15356	11850	14400	13605	11829	16330
Quadratic fit	4306	6517	4432	5094	5857	5028	5193	5380	5130	5769
Exponential fit	5006	7977	4921	6022	6898	5808	7749	6422	5838	8491
Logarithmic fit	18192	32453	19657	23767	22490	15664	22399	18987	15454	25735
Power Regression	11328	19351	12122	14393	15007	11730	13806	13216	11650	15683
LNP fit	29150	59576	32446	41329	39075	22463	37512	31025	22324	45048

Table 5.23 – Average Absolute Error (Rate-SSIM; Picture Type; GOP Pattern)

Fit Method	IPPP GOP1		IPPP GOP2		IBBP GOP1			IBBP GOP2		
	I Type	P Type	I Type	P Type	I Type	P Type	B Type	I Type	P Type	B Type
Linear fit	0.64	0.37	1.03	0.34	1.81	0.60	0.41	1.11	0.65	0.41
Quadratic fit	0.25	0.14	0.41	0.13	0.68	0.22	0.16	0.41	0.23	0.15
Exponential fit	1.86	1.08	3.03	0.99	5.06	1.69	1.14	3.12	1.84	1.13
Logarithmic fit	0.91	0.53	1.44	0.48	2.54	0.85	0.56	1.59	0.94	0.57
Power Regression	2.45	1.42	4.00	1.31	6.68	2.23	1.51	4.12	2.42	1.49
LNP fit	4.16	2.44	6.63	2.22	11.39	3.82	2.55	7.15	4.20	2.55

Table 5.24 – Average Absolute Error (SSIM-QP, Picture Type; GOP Pattern)

Table 5.23 and Table 5.24 present the combined results of the fit methods regarding the different GOP patterns and picture type. Individual results per video sequence are available at the Annex B according to the distinct fit methods, GOP pattern and picture type. Regarding Rate SSIM, average results from Table 5.23 point to quadratic approach as the best method. Good results are obtained with the exponential method. In fact, for one in each four sequences exponential method performs better than quadratic function. This occurs in sequences with average spatial complexity and high temporal complexity or high spatial complexity and average temporal complexity.

As for SSIM-QP, the results converge into one function: quadratic function. Quadratic function obtained the lowest value for all the individual sequences, regardless of picture type or GOP Pattern. For SSIM-QP, the second best choice is linear fit. This is an interesting result. Linear approach is the least complex model of the selected set of function and is widely used when computational power is a limited resource.

From Table 5.25 to Table 5.28, it can be observe R-D and D-QP results for Akiyo, Foreman and Football, for the different GOP Patterns. These examples illustrate the impact of the different approaches on the absolute error, particularly in Rate-SSIM, regarding picture type, GOP pattern and content complexity.

Sequence	Fit Method	SSIM-QP		Rate - SSIM	
		I Type	P Type	I Type	P Type
Akiyo	Linear fit	0.66	0.39	2860	9658
	Quadratic fit	0.08	0.05	1303	3705
	Exponential fit	1.15	0.67	1760	5025
	Logarithmic fit	1.02	0.60	3348	11833
	Power Regression	1.51	0.88	2420	7778
	LNP fit	3.37	1.96	6771	27294
Foreman	Linear fit	0.37	0.20	8973	15467
	Quadratic fit	0.21	0.12	2815	4171
	Exponential fit	1.58	0.94	3053	7355
	Logarithmic fit	0.55	0.33	12730	27411
	Power Regression	2.16	1.27	9211	15722
	LNP fit	3.78	2.22	20668	49943
Football	Linear fit	1.01	0.55	11442	13959
	Quadratic fit	0.44	0.25	4043	4858
	Exponential fit	2.03	1.16	10007	13079
	Logarithmic fit	0.70	0.37	24529	33171
	Power Regression	2.86	1.65	12886	16915
	LNP fit	3.92	2.32	44254	62171

Table 5.25 – SSIM-QP and Rate-SSIM average absolute error for Akiyo, Foreman, Football (IPPP GOP1)

Sequence	Fit Method	SSIM-QP		Rate - SSIM	
		I Type	P Type	I Type	P Type
Akiyo	Linear fit	1.04	0.35	3720	5806
	Quadratic fit	0.12	0.04	1567	2291
	Exponential fit	1.82	0.61	2062	3109
	Logarithmic fit	1.61	0.54	4455	7059
	Power Regression	2.39	0.80	3024	4713
	LNP fit	5.35	1.79	9680	15955
Foreman	Linear fit	0.66	0.19	9344	11398
	Quadratic fit	0.36	0.12	2731	3339
	Exponential fit	2.47	0.86	2583	4804
	Logarithmic fit	0.85	0.29	13684	18684
	Power Regression	3.42	1.16	9332	11754
	LNP fit	5.93	2.02	22778	32713
Football	Linear fit	1.71	0.55	11815	12077
	Quadratic fit	0.70	0.23	4425	4291
	Exponential fit	3.15	1.05	9932	10930
	Logarithmic fit	1.19	0.38	25615	26846
	Power Regression	4.51	1.49	13410	13906
	LNP fit	6.11	2.06	46990	49344

Table 5.26 – SSIM-QP and Rate-SSIM average absolute error for Akiyo, Foreman, Football (IPPP GOP2)

Sequence	Fit Method	SSIM-QP			Rate - SSIM		
		I Type	P Type	B Type	I Type	P Type	B Type
Akiyo	Linear fit	1.86	0.62	0.42	5491	1343	4664
	Quadratic fit	0.18	0.06	0.04	2177	772	1992
	Exponential fit	3.16	1.06	0.72	2943	1001	2636
	Logarithmic fit	2.84	0.95	0.64	6685	1476	5556
	Power Regression	4.13	1.38	0.94	4428	1226	3833
	LNP fit	9.19	3.08	2.08	15265	2280	11719
Foreman	Linear fit	1.13	0.37	0.26	8743	6943	10749
	Quadratic fit	0.59	0.23	0.17	2923	3100	3618
	Exponential fit	3.84	1.45	1.08	3439	3667	4637
	Logarithmic fit	1.34	0.51	0.37	11966	8698	16600
	Power Regression	5.40	1.99	1.45	7597	7111	10961
	LNP fit	9.85	3.46	2.40	23804	13009	27620
Football	Linear fit	3.23	1.01	0.76	12107	10684	12339
	Quadratic fit	1.11	0.39	0.32	4235	3425	5692
	Exponential fit	4.77	1.76	1.33	12779	10666	13899
	Logarithmic fit	2.19	0.68	0.53	28060	21627	29266
	Power Regression	7.08	2.54	1.90	12973	10083	15190
	LNP fit	10.29	3.55	2.37	53589	38445	54153

Table 5.27 – SSIM-QP and Rate-SSIM average absolute error for Akiyo, Foreman, Football (IBBP GOP1)

Sequence	Fit Method	PSNR-QP			Rate - PSNR		
		I Type	P Type	B Type	I Type	P Type	B Type
Akiyo	Linear fit	1.19	0.70	0.43	3676	1277	6564
	Quadratic fit	0.11	0.06	0.04	1543	734	2651
	Exponential fit	2.00	1.17	0.73	2046	954	3601
	Logarithmic fit	1.81	1.05	0.65	4432	1402	7932
	Power Regression	2.62	1.52	0.94	2994	1167	5377
	LNP fit	5.80	3.37	2.08	9881	2161	17445
Foreman	Linear fit	0.61	0.35	0.23	7966	6745	12567
	Quadratic fit	0.40	0.21	0.15	3078	3037	3987
	Exponential fit	2.54	1.56	1.07	3514	3786	5246
	Logarithmic fit	0.94	0.56	0.39	10417	8392	20353
	Power Regression	3.50	2.13	1.43	7515	6922	12925
	LNP fit	6.43	3.81	2.42	18621	12534	34639
Football	Linear fit	1.84	1.03	0.72	11650	11197	13148
	Quadratic fit	0.73	0.45	0.33	3928	3630	5946
	Exponential fit	3.30	2.03	1.39	10797	9568	13794
	Logarithmic fit	1.26	0.70	0.50	24457	21847	31480
	Power Regression	4.73	2.88	1.95	11741	10509	16693
	LNP fit	6.68	4.03	2.42	44929	38346	57902

Table 5.28 – SSIM-QP and Rate-SSIM average absolute error for Akiyo, Foreman, Football (IBBP GOP2)

5.4 Bit Rate Variability-Distortion for H.264/AVC

The broadcast of video streams over networks has become a popular service in the last years. In order to ensure high network utilization, the video streams are typically transported with some sort of statistical transport scheme ([447]). According to S. K. Srinivasan et al. ([448]) statistical multiplexing of encoded video sequences can be performed with or without coordinating the encoders of the multiplexed streams. On the one hand, statistical multiplexing is performed using interconnected video encoders. In this case, encoding video parameters of the individual streams are altered such that the combined video traffic conforms with the available network bandwidth. On the other hand, no coordination between encoders is performed and thus encoding parameters are kept constant (for example, frames may be dropped). Several studies have focused on the analysis and modeling of video traffic and video network transport mechanisms ([449],[450]). The study of the video encoder's statistical characteristics and compression performance from a communication network perspective has received considerable attention ([447],[451],[452],[453],[454]). One of the reasons is the potential to improve the efficiency of video transport over communication networks using statistical multiplexing ([447],[451]). In general, these studies focused on VBR encoded video with fixed QP. In this section, the relationship between the variability of the bit rate of the video sequences will be analysed as a function of the quality level (PSNR) when video is encoded in an open loop with a fixed quantization scale. Patrick Seeling and Martin Reisslein [451] introduced the bit rate variability-distortion (VD) curve. One of the findings was that the VD curve exhibits a characteristic "hump" behavior. The VD curve relates the bit rate variability of an encoded video sequence to its average quality level measured by PSNR. The bit rate variability is usually characterised by the Coefficient of Variation (CoV) of the size of the frame (in bits). CoV is defined as the ratio between the standard deviation of the frame sizes normalized by their mean ([433],[434]):

$$CoV = \frac{\sigma}{\bar{X}} \tag{5.50}$$

where $\bar{X}$ is mean of the frames size (in bits), M the number of frames

$$\bar{X} = \frac{1}{M}\sum_{m=1}^{M} X_m \tag{5.51}$$

and the variance σ^2 (square of the standard deviation) of the frame sizes being defined as

$$\sigma^2 = \frac{1}{(M-1)}\sum_{m=1}^{M}\left(X_m - \bar{X}\right)^2 \tag{5.52}$$

It is also possible to find in the literature, the peak-to-mean frame size ratio, that is, the ratio between the largest encoded frame and the average encoded frame size ([451]). In contrast to the existing studies, we also investigate the impact of replacing PSNR by perceptual quality metrics in the VD curve. Simulations were conducted using 12 video sequences, open-loop coding setup, and QP ranging from 10 to 40. The H.264/AVC JM reference software encoder, version 11.0, was used in the Main Profile, CABAC, Hardamard On, RDO On, with two reference frames from the past and the future (Table 5.2).

5.4.1 *Bit Rate Variability as a function of PSNR*

Results were grouped into two video sub-sets. The first video subset, left side of Figure 5.12, contains video sequences with a medium to high level of spatial detail or temporal complexity (Foreman, Football, Coastguard, Flower Garden, and Mobile and Calendar). The second video subset, right side of Figure 5.12, contains video sequences with a fixed camera and low to medium spatial detail and motion activity (Akiyo, Deadline, Hall, Mother and Daughter, News, Paris, and Silence). Analysing the R-D charts, video sequences encoded with B slices achieved higher RD efficiency than video sequences encoded only with I and P slices. Nevertheless, the substantial increment in compression efficiency with B slices occurred at the expense of increased video traffic variability, as indicated by the significantly higher CoV values. For example, the maximum value of CoV in sequences with B slices (2.9) occurred with IBBP GOP1 pattern and the minimum value occurred with (approximately 1.5). This GOP structure corresponds to video sequences with the highest and lowest number of Intra coded slices, respectively. The cause for this great difference in bit rate variability is the enhanced compression performance of the H.264/AVC encoder. Particularly, the advanced motion compensated prediction results compared with the enhancement of the I-frame because of the spatial intra prediction. Therefore, the mixture of all the new compression tools is responsible for the higher bit rate variability for the H.264/AVC encoder. A sharp drop of the bit rate variability for the sequences encoded with B slices is also observed. Bit rate variability is lower for sequences containing high spatial detail or motion activity than for sequences with lower spatial and temporal complexity. In the next section this analysis will be repeated for different quality metrics.

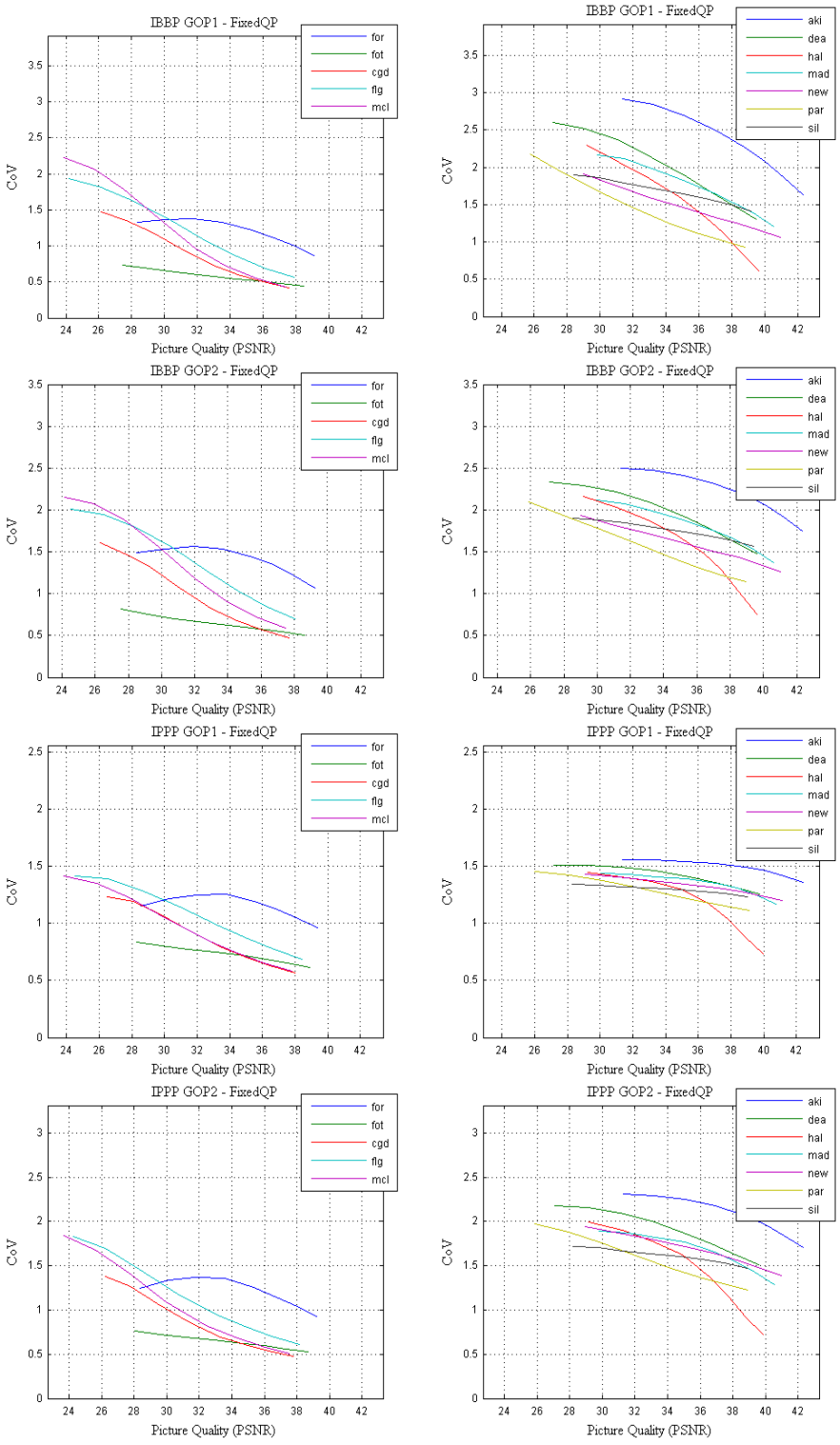


Figure 5.12 – Rate Variability-distortion (VD) Curves (PSNR)

5.4.2 *Bit Rate Variability as a function of Perceptual Metrics*

We have modelled the bit rate variability-distortion (VD) curves for the twelve video sequences and the different GOP patterns (Figure 5.13 to Figure 5.15). Instead of using PSNR as the quality metric, metrics based on JND (PSPNR and SAD_{JND}, and SSD_{JND}.) and SSIM metrics were selected. The PSPNR and SSIM metrics assess picture quality and the remaining two perceptual distortions. To provide a better reading of the results, they were aggregated into two video subsets according to spatial and temporal complexity. One subset composed of the video sequences Foreman, Football, Coastguard, Flower Garden, and Mobile and Calendar. Another sub-set composed of the video sequences Akiyo, Deadline, Hall, Mother and Daughter, News, Paris, and Silence.

There is an inversely proportional relationship between PSPNR and SAD_{JND}, SSD_{JND}. This can be observed directly in the different charts. For higher values of PSPNR, the variability decreases whereas for higher values of SAD_{JND}, SSD_{JND}.the variability increases. Video sequences with lower spatial-temporal complexity present higher values of bit rate complexity. This is in line with H.264/AVC new prediction tools: inter prediction is more effective and thus requires a lower number of bits. Again, sequences with better RD performance present higher values of bit rate variability. Similar behaviour is observed in relation to the GOP pattern influence on bit rate variability. Sequences encoded with B slices present much higher values of CoV. Longer size GOP present higher values of bit rate variability.

It can be noted that SSD_{JND}.VD curves present a very sharp variation for video sequences with low complexity up to a certain value. After this threshold, the level of bit rate variability increases at a lower rate. This phenomenon is also observed in VD SAD_{JND}, curves but it is not as severe. VD curves for PSPNR and SSIM are, in general, smoother compared with VD curves for the distortion metrics.

Results vary with content nature and GOP pattern. When sequences are more complex to encode, the relationship between texture bits and motions bits is more balanced and bit rate variability decreases. Nevertheless, H.264/AVC global results indicate high bit rate variability when using perceptual image quality metrics. This is a good result for statistical multiplexing. Nevertheless, it should be noted that the peak of CoV varies greatly with content and coding parameters.

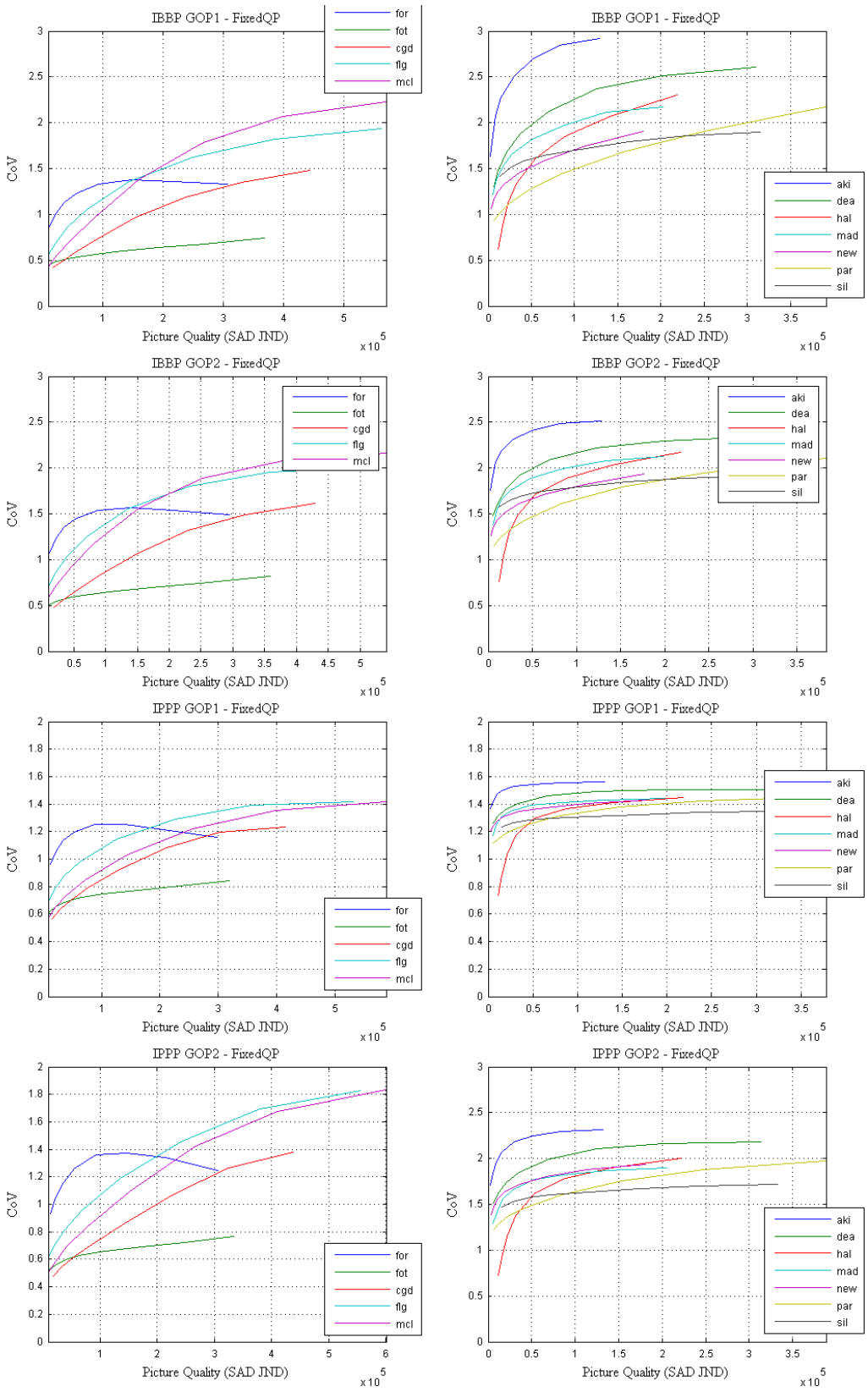


Figure 5.13 – Rate Variability-distortion (VD) Curves (SAD_JND)

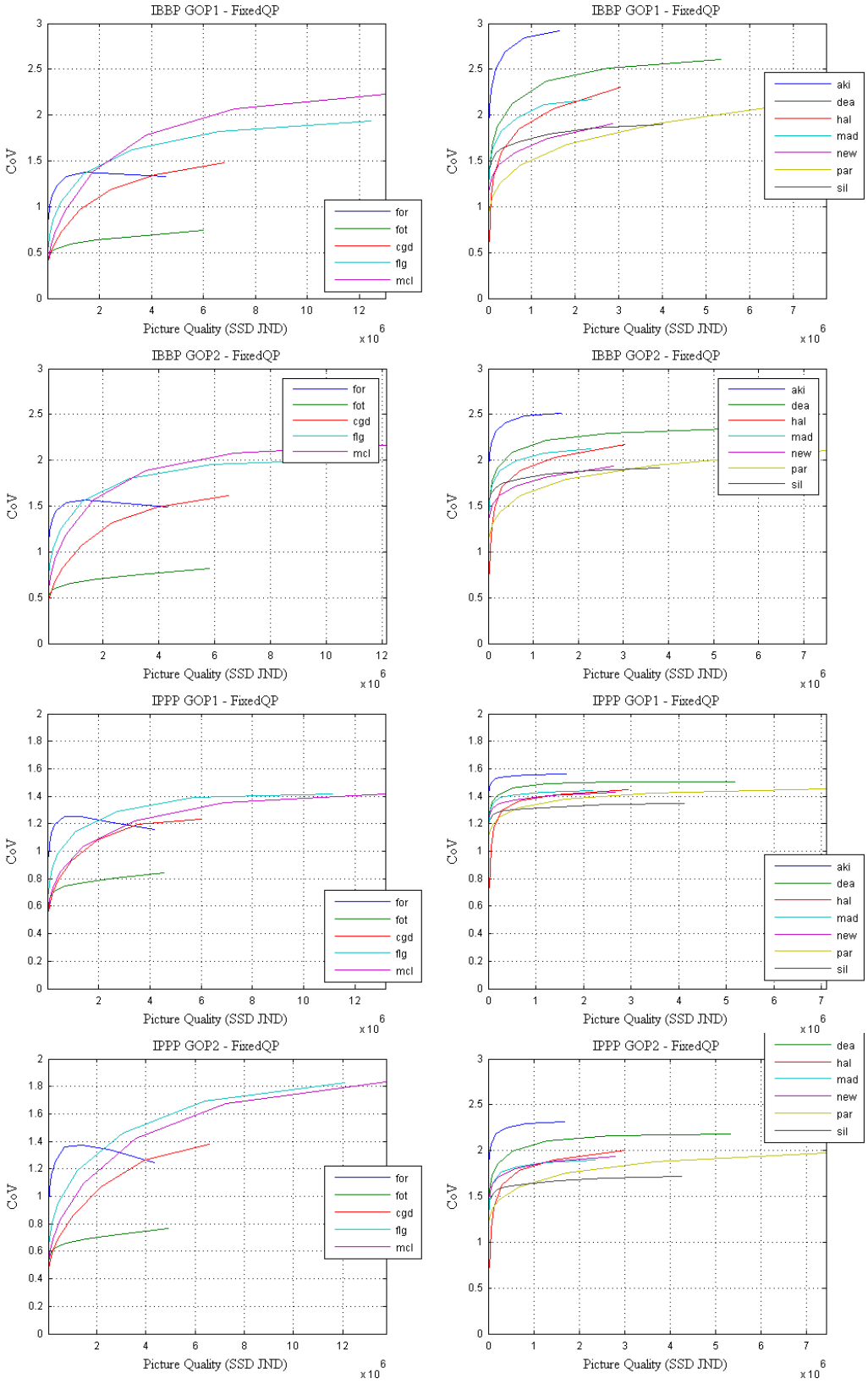


Figure 5.14 – Rate Variability-distortion (VD) Curves (SSD_JND)

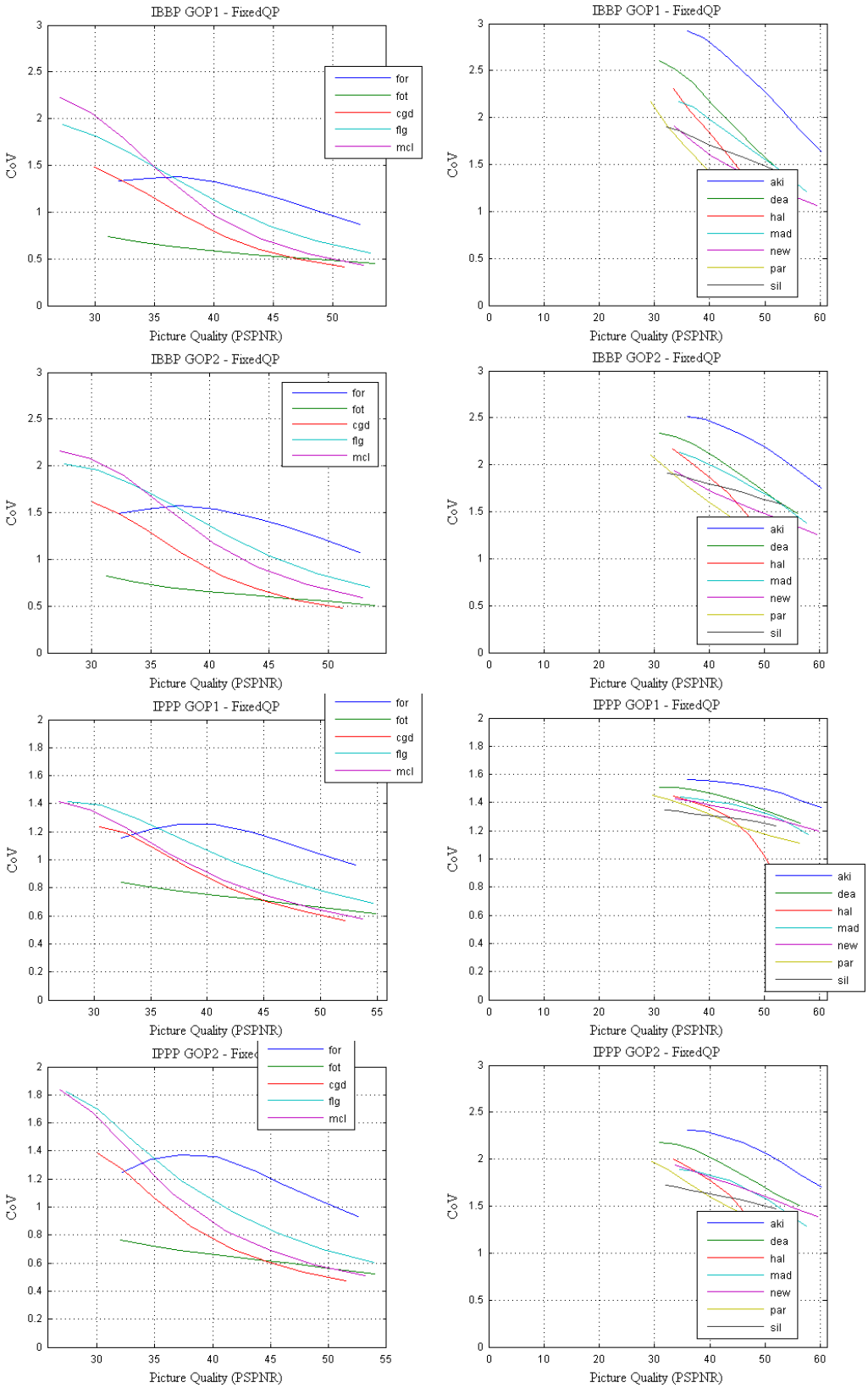


Figure 5.15 – Rate Variability-distortion (VD) Curves (PSPNR)

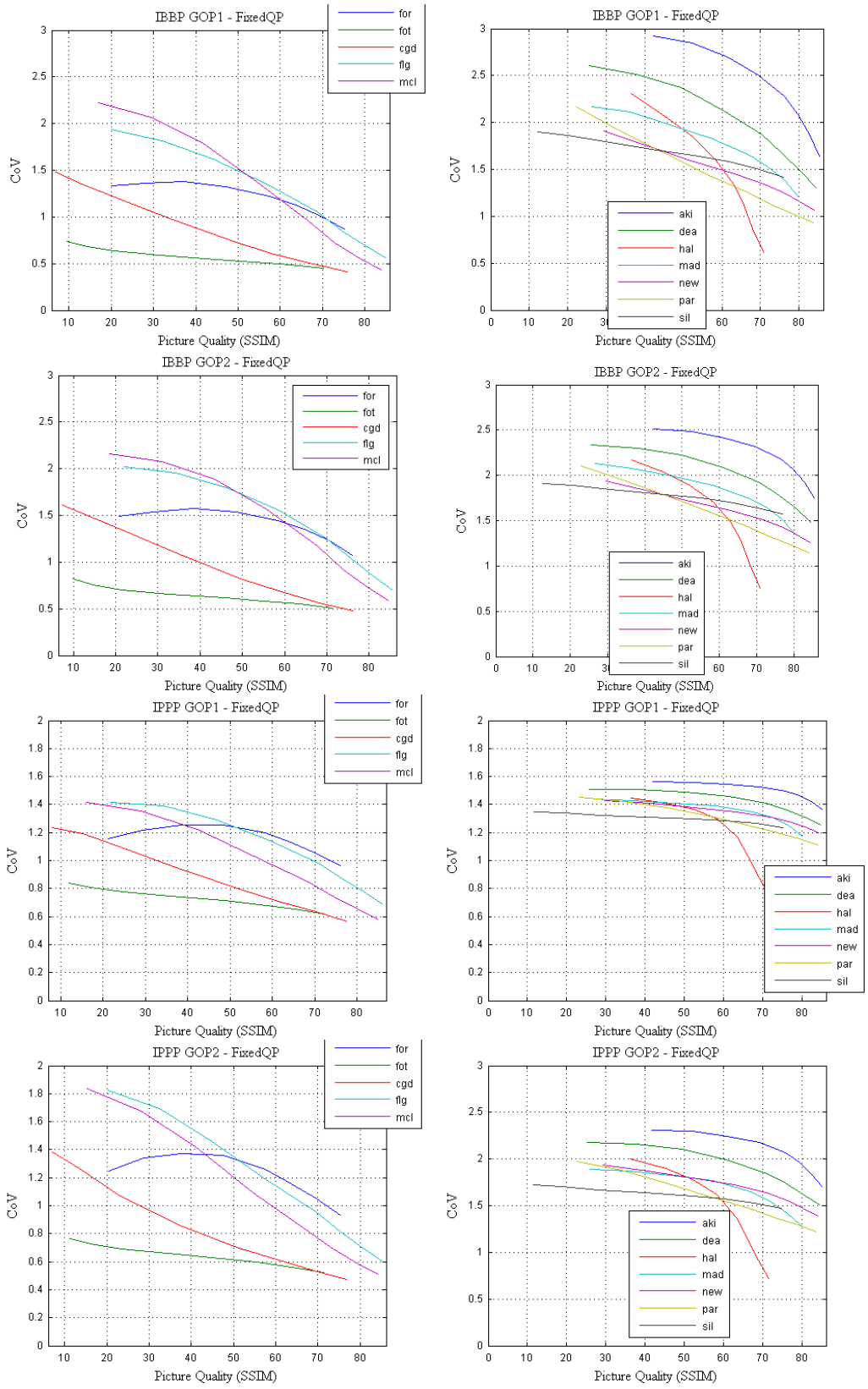


Figure 5.16 – Rate Variability-distortion (VD) Curves (SSIM)

5.5 Summary

As was mentioned at the beginning of this Chapter, the main goal of rate control is to present high-quality video to end users by setting the proper coding configurations within constrained bounds such that the buffer does not overflow or underflow, at limited storage capacity or channel bandwidth ([18],[331]). Rate-distortion (R-D) based methods are often implemented to enhance and stabilize video quality. The MSE and PSNR are the most used video quality metrics, given their little computational complexity, regardless of their limitations. In fact, MSE and PSNR do not correlate well with perceived quality. While MSE measures the image difference, PSNR measures image fidelity. A number of alternatives have been proposed in recent years. Two different FR approaches for quality assessment, based on JND and on structural similarity (SSIM), have had their concept, implementation and meaning presented in Chapter 2. In contrast with traditional metrics, they use mechanisms to incorporate HVS or the perceptual effects of video degradation. As a result, they allow a more refined prediction of the level of degradation that a signal can suffer until a human observer notices it ([212]).

In this chapter, previous research conducted in the field of Rate Control optimization was examined, particularly R-D modeling and the developing of R-D functions for the rate control of joint video sequences using perceptual quality metrics. In particular, extensive experiments on a large number of video sequences were performed, their statistics studied and a Rate-Quantization (R-Q) model and Distortion-Quantization (D-Q) model derived for modeling the R-D relation in H.264/AVC.

Experimental results show that quadratic function is a good solution in most cases for SSIM and PSPNR. In most cases, quadratic approach is the best solution regarding Rate and SSIM and Rate and PSPNR, while logarithmic and power methods are better for SAD_{JND} and SSD_{JND} metrics. Results have been verified for different GOP Patterns, video test sequences, and coding setup. By using perceptual models, average perceptual quality improvement can be achieved when the proposed model is exploited in predicting the rate of the rate-control scheme. The proposed models can be applied to Intra and Inter prediction frames. In the last section, a bit rate variability study was conducted in order to assess the level of variability of an H.264/AVC video stream. High values of CoV were observed particularly for video sequences encoded with B slices. Based on improved R-D models, a novel proposal for jointly coding of multiple video sequences will be proposed using perceptual image quality metrics, implemented into the H.264 reference software JM reference software version 11.0 and studied in Chapter 6.

Chapter 6. Joint Video Encoding of H.264/AVC bitstreams

This chapter examines how several encoders can operate together to enhance global picture quality. This theme is an extension of rate control of the independent video encoding of several programmes. First, the concepts of statistical multiplexing and joint video coding are introduced followed by a review of current systems available in the literature ([280],[300],[455],[456], [457],[458]). Their limitations are addressed, and potential improvements to the algorithms are discussed. Secondly, joint coding algorithms are presented and assessed using objective and subjective image quality metrics. They attempt to allocate the existing bandwidth, according to the coding complexities of video sources, measured by perceptual metrics, and to uniformize the level of perceptual quality of the encoded video. Each encoder uses an independent rate controller to allocate bit rate within each picture. The algorithms work with a look ahead window of one GOP size. Results point to a reduction in the amplitude of quality variation among programmes compared with the CBR scenario and a decrease of the bit rate variability along the bitstreams. Subjective evaluation was conducted with a panel of 15 viewers and using SAMVIQ methodology. Thirdly, a two-pass encoding strategy that incorporates perceptual information is presented and discussed.

6.1 Statistical Multiplexing and Joint Video Encoding

When H.264/AVC encoders work in CBR mode, the picture quality varies depending on the complexity of the video signal. Usually, viewers assess picture quality based on the pictures with the highest amount of impairment. Thus, to guarantee that even the most critical pictures are encoded at an acceptable quality level, the bit rate should be set at a high value. However, for most of the duration of a video programme, a lower bit rate would be entirely sufficient. In order to generate a constantly high picture quality, the bit rate should be allowed to vary in accordance with the complexity of the video signal. For example, the DVD specification supports MPEG-2 video encoding, where the bit rate changes from scene to scene from about 1 Mbps up to 9.8 Mbps, at an average video bit rate from 3.5 up to 6 Mbps. An equivalent idea can be applied when several programmes are encoded in VBR mode and transmitted

simultaneously by sharing the total available bandwidth between them. Thus, critical video signals requiring higher bit rates can obtain a larger part of the total bandwidth than non-critical video signals. In general, bit rate is allocated among programmes according to the statistics of the video signals, and so these systems are named ‘statistical multiplexing systems’ (also designated by StatMux) ([280],[300],[455],[456],[457]). Usually, when the number of channels is reduced, picture quality can be increased by preventing the worst impairments during difficult scenes. When the number of channels is larger, besides an increment in picture quality, these systems can also make the reduction of the overall bandwidth possible.

In statistical multiplexing systems, a constant bandwidth communication channel is virtually segmented into different variable bandwidth channels. The channel's bandwidth can be altered according to the instantaneous traffic requirements of the programmes that are being sent over the channel. An increment in the allowed variation in the bit rate corresponds to an increment in the initial buffering delay. H.264/AVC supports the use of larger buffers than previous standards, such as MPEG-2, so that it can absorb momentary peaks in bit rate, during a short period of time, without altering QP. Larger buffer size corresponds to a greater end-to-end delay. Although this value is minimised when programmes are encoded in CBR mode, a moderate delay lets full advantage be taken of encoding in VBR mode. To control bit rate variations a joint rate control algorithm may be used. Joint video encoding is a special case of statistical multiplexing systems. In this scenario, statistical multiplexing is performed in conjunction with encoding so that a common bit budget is divided between the bitstreams, giving their temporal complexities.

Statistical multiplexing systems differ from single VBR channel's capacity. A VBR channel, for short periods of time, can absorb peak bit rate requests. Nevertheless, the scene duration may last longer than the channel capacity to absorb bit rate variations. In addition, during network congestion, this capacity is further reduced and may result in a decrease of picture quality. In a StatMux system, the goal is to reallocate bandwidth as and where it is needed. This concept differs from the initial example of the DVD. When video is encoded for storage, the rate control algorithm decides when and how to vary bit rate to obtain uniform picture quality and thus performs a kind of statistical multiplexing in the time domain. In the case of StatMux, the multiplexing is performed across numerous video channels.

The different time scales regarding bit rate control within a programme should be noted. For example, in Europe, a picture is encoded in 40 ms, and a macroblock from about 100μs (CIF) up to 5μs (HDTV) ([300]). Usually, a GOP structure corresponds to 500 ms. Within a programme, temporal activity, either from a camera movement (such as a zoom or a pan), video editing (such as a transition or a special effect) or content itself, can last from a few seconds to

several minutes. Depending on the programme nature (talks, sports, cinema, etc.), a scene change can occur in a couple of seconds (sports) up to many minutes (cinema). The rate control algorithm of an encoder should focus on short-term variations resulting from macroblock coding up to picture coding, whereas a joint rate controller should operate on the longer time scale. In most cases, existing joint rate controller schemes operate based on the time scale corresponding to a GOP structure. As discussed in the previous chapter, GOP pattern is associated with the minimum decoder refresh delay to enable channel switching. Preceding standards of H.264/AVC used GOP structure to define random access but also to avoid decoder drift as DCT and inverse DCT are defined with only a certain accuracy. As referred in Chapter 3, H.264/AVC defines a forward and inverse transformation with full accuracy so this problem is solved. Thus, the coding of an intra-slice picture is associated with random access requirements. Whenever numerous programmes are transmitted at the same time two main approaches should be considered: independent video encoding and joint video encoding. The major difference is whether or not programmes share a common bit budget and the way this budget is shared. The next section will discuss both scenarios.

6.1.1 *Independent Video Encoding of Multiple Programmes*

Each programme in this scenario is encoded independently without using any information from the remaining programmes.

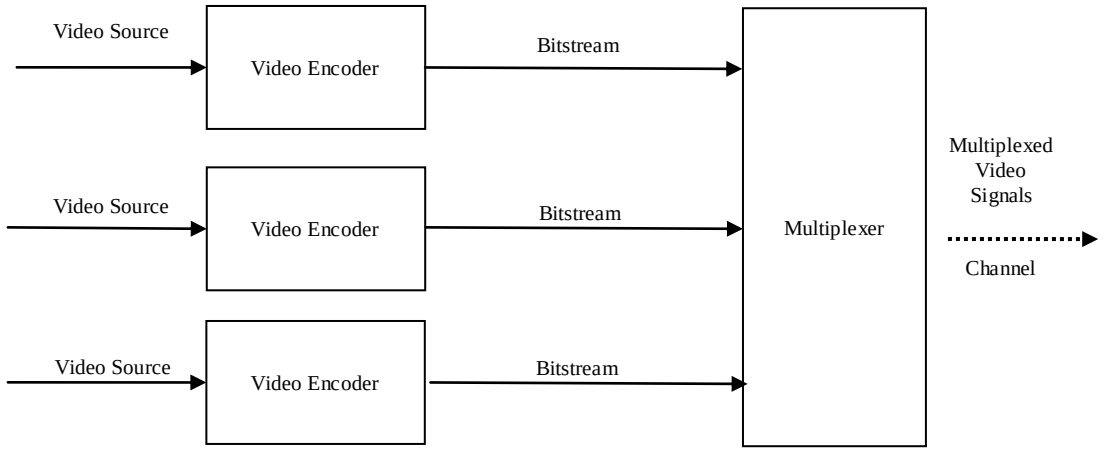


Figure 6.1 – Block Diagram of Independent Video Coding

Figure 6.1 displays a block diagram for this scenario. Each video encoder has a separate rate controller. After encoded the video source the bitstream is sent to the Multiplexer. A programme can be encoded in CBR or VBR mode. If the programmes are encoded at CBR, the multiplexed video stream will occupy a bandwidth that corresponds to the sum of the individual bandwidth

of the programmes. This method is a simple process and is known as deterministic multiplexing. The major drawback is that a large amount of bandwidth is wasted for bursty programmes. This is special relevant as H.264/AVC due to its compression efficiency, generates traffic, which is extremely burst over a wide range of time scales ([447],[449],[450],[451],[452],[453], [454],[459],[460]). This is not an efficient way to use resources. If the programmes are encoded in VBR mode, then instead of the average, one solution is to use the peak bit rate of each programme that needs to be used. An alternative to deterministic multiplexing is to use statistical multiplexing. In this case, a statistical multiplexing gain is obtained by allocating to the multiple programme's streams a lower bandwidth value compared with the sum of the peak bit rates. Therefore, more programmes may be broadcasted and thereby a better utilization of the resources is achieved. Although the instantaneous joint bandwidth resulting from multiplexing all the different programmes may exceed the channel capacity, the aggregated bandwidth may generate smoother video traffic.

Several VBR rate control strategies have been presented in the literature. One possible solution is to maintain the QP constant. This strategy is also known as open-loop ([51],[280],[455]). The goal is to obtain uniform quality level by introducing the same level of distortion in the encoded video stream. In this scenario, the encoder may use as many bits as needed to achieve a predetermined quality level. Thus, when the rates of all the VBR encoders are combined, they may exceed the channel capacity. In particular, if the burst of bits take place at the same time. In this case, the buffer will overflow and data will be lost. With open-loop VBR it is hard to achieve both good channel utilization and very limited data loss ([280]). Another approach is to control the variation of VBR in order to obtain constant quality (CQ-VBR) ([51]) and to guarantee that the buffering delay does not exceed a pre-define threshold value. This approach is also designated in the literature as close-loop ([280],[455]). In this scenario, feedback information from the encoder process regarding the video source is used in the rate control algorithm. Feedback may be used according to different approaches: "feed forward rate control" and "feed backward rate control" (Figure 6.2). In the feed backward, there is limited knowledge of the sequence complexity. The encoders gather statistical information during the encoding process. This information can be used to determine the video complexity of the programme. Bits are allocated on a picture basis and spatially uniform distributed throughout the image. This approach assumes that neighbouring frames share an equivalent coding characteristic. Thus, this approach often suffers from performance degradation at scene changes.

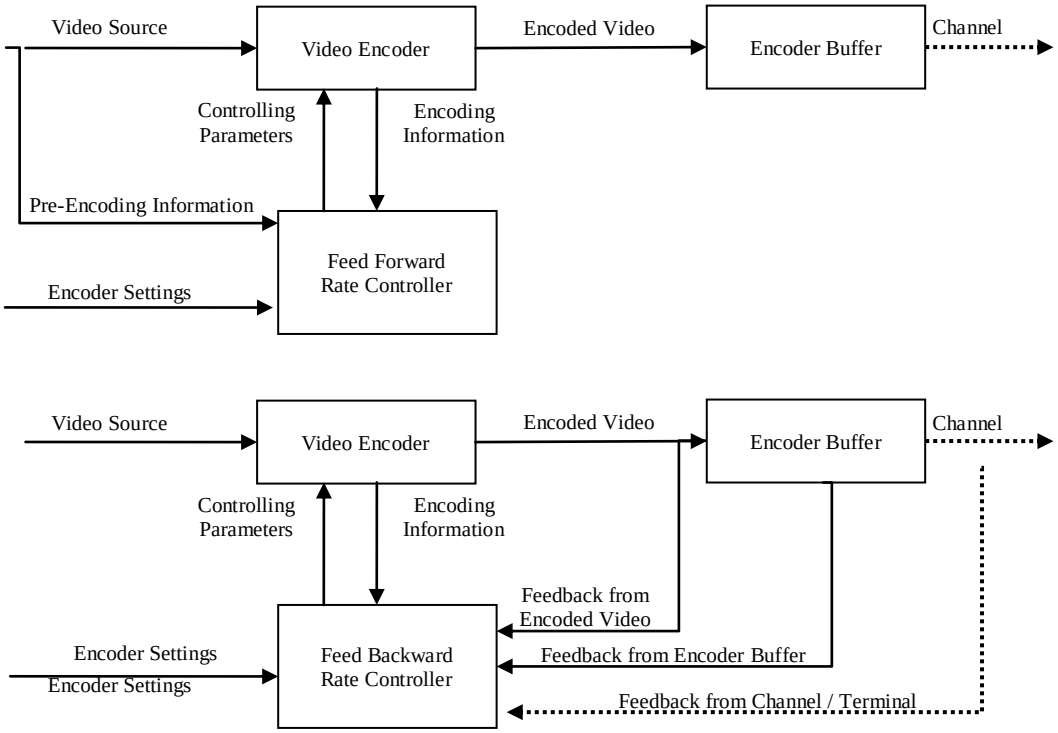


Figure 6.2 – Block Diagram for Feed-Forward and Feed-Backward rate control

In the "feed forward rate control," a pre-analysis is performed on “future” frames to be encoded in order to collect R-D statistics. The collected statistic's information is used for rate allocation. A good R-D model for video encoders with a mathematical framework for joint coding allows to obtain an increase in performance. For off-line applications, without time constraint, two-pass VBR rate control techniques are often used to optimize video quality. In this case, part or the entire video sequence is first encoded using a fixed quantisation step size in order to extract R-D statistics ([461],[462],[463],[464],[465],[466],[467]). In the second-pass, the quantisation parameters are adjusted, in order to obtain stable video quality or superior coding performance. In one of the early examples ([461],[462],[463],[464],[465]), the author proposes a two-pass rate control strategy based on obtain the Bit Usage Profile. The pre-analyser (first step) generates a histogram of the number of bits, the quantisation parameter and the PSNR for each macroblock (both for a particular position and regardless of the spatial position). The Bit Usage Profile allows determining how, when and where the complexity varies. Thus, in the second-pass, the result will present an improvement of the perceptual picture quality ([462],[463]).

Modeling VBR video programmes have been extensively studied in the literature. Proposals include first-order auto regressive (AR) models, discrete AR (DAR) models, Markov renewal processes (MRP), MRP transform-expand-sample, finite-state Markov chain, Gamma-beta-auto-regression (GBAR) models, discrete-time semi-Markov processes (SMP), wavelets, multifractal

and fractal methods, Log-normal, Gamma, and hybrid Gamma/ Lognormal distribution model ([468],[469]). Recently, Aggellos in [468] proposed the simple Discrete Autoregressive model (separately for I, B, and P frames) to capture the behaviour of multiplexed H.264/AVC video conference sources. These models are obtained by analysing the bit rate characteristics of the encoded programmes (the histogram, in general, converge towards a probability density function (PDF)). An example of a histogram and the corresponding Gamma PDF is presented in Figure 6.3 ([51]).

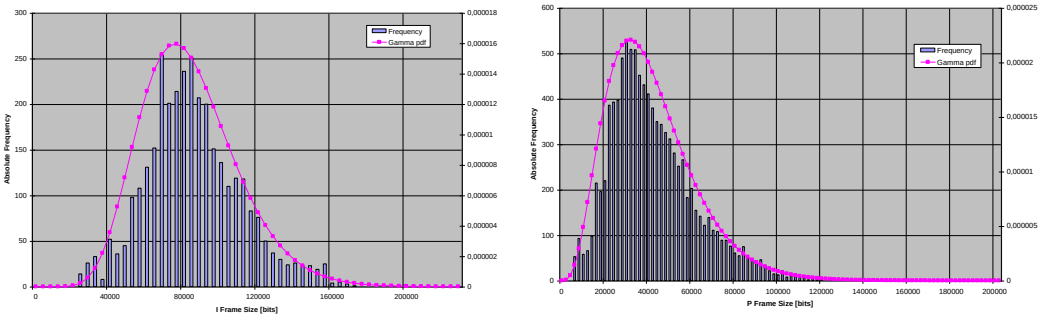


Figure 6.3 – Histogram of bit rate for Bond sequence

When several programmes are statistically multiplexed, their combined bit rate PDF can be obtained by convolution of the individual PDFs. To determine the statistical multiplexing gain first it is computed the cumulative distribution function (CDF). Bit rate is selected depending on the requirements of picture quality. The major drawback in this approach is that the bit rate in H.264/AVC suffers from a high variability, and long programmes are not stationary in time ([448],[452],[453],[460]). According to Auwera et al. this variability is magnified in H.264/SVC due to the improved number of motion estimation modes (H.264/AVC) and hierarchical B-frames (H.264/SVC). In the paper, different levels of smoothing that give (bufferless) statistical multiplexing performance close to an optimal off-line smoothing technique are identified. Nevertheless, the method still lacks the integration of collaborative smoothing strategies with active buffer managements. Geert and Reisslein in [460] examined the statistical multiplexing behaviour of H.264/SVC, H.264/AVC, and MPEG-4 Visual with long video sequences. The levels of smoothing that give (bufferless) statistical multiplexing performance close to an optimal off-line smoothing technique and the size of the multiplexer buffer are identified. The next steps will promote the integration of collaborative smoothing strategies with active buffer managements. Mehdi proposed a StatMux technique in conjunction with the time-slicing transmission scheme by the IP encapsulator in a DVB-H network ([470],[471],[472]). This method result in a decrease of the end-to-end delay of DVB-H services compared with a deterministic approach. Vladimir Vukadinovic et al. in [450] examine the potential statistical multiplexing gains with and without the coordination in the encoders, in

Multimedia Broadcast and Multicast Service (MBMS) and enhanced MBMS (E-MBMS) of 3GPP. E-MBMS supports the mapping of multiple MBMS services on the same multicast channel (MCH) ([473]). In [473] one mobile TV channel is reported to be around 300kbps (H.264/MPEG4), and the audio encoder outputs 32~64kbps stream according to different configurations (AAC). Results also showed that in the case of coordinated encoding, channel allocation updates do not contribute significantly to the gains in terms of average PSNR/bit rate, but may reduce the PSNR variations within a stream and thus provide gains in terms of visual quality.

Cheng-Hsin Hsu et al. studied the problem of broadcasting multiple VBR programmes over a broadcast network to many mobile devices, in close-loop and open-loop ([474],[475],[476],[477]). In close-loop, each video stream is controlled by a joint video coder. In open-loop, there is no joint video coder and thus video streams may occasionally overload the broadcast network. Two performance metrics for video streaming over wireless networks have been defined: energy saving and goodput, from mobile users' and network operators' point of view, respectively. The problem has been mathematically formulated as a burst scheduling problem for multiple TV channels with arbitrary bit rates. To solve this problem, Cheng-Hsin Hsu et al. proposed an algorithm to perform the scheduling of the programmes: the Statistical Multiplexing Scheduling (SMS). Experimental and simulation outcomes show that the resulting schedule performs well, and that results are inline with most practical networks.

Martin Fleury et al. ([478],[479]) proposed a system that combines a bank of bit rate transcoders and a statistical bandwidth manager. The system uses two metrics to measure the content complexity: the temporal complexity index (TI) and the Scene Complexity Index (SCI). Using a fuzzy logic controller the metrics are combined. Both metrics are determined across a GOP. Experimental results were obtained using the H.264/AVC JM reference software version 14.2, RDO ON, and three CIF video streams (30 fps). Each of three video streams contained 900 frames consisting of FNS (Foreman + News + Stefan), NMF (News +Mobile + Foreman) and WHB (Flower + Highway + Bus). An IPPP... GOP structure was set with Instantaneous Decoder Refresh (IDR) frames configured, and the intra frame refresh was set to 15. Results show a reduction in quality fluctuations compared with independent video coding. Nevertheless, global results show a decrease in overall picture quality.

Mehdi proposed a fuzzy joint encoding and statistical multiplexing scheme for streaming over DVB-H ([472],[480],[481]). The goal was to decrease the end-to-end delay in a broadcast system by decreasing the buffering delays. The rate control algorithm uses several fuzzy controllers to control the bit rate of each encoded bitstream and also the bit rate of the aggregated bitstream. Simulation's results were conducted for a group of four video sequences,

with duration of 60 seconds, a frame rate of 15fps, QVGA, and a target bit rate of 300kbps. The proposed algorithm provided a 38% reduction in the required decoder buffer size and 62% reduction in the decoder buffering delay at the expense of 0.02dB degradation in quality. Nevertheless, these results are obtained without reducing the variations in picture quality between different programmes.

6.1.2 *Joint Video Encoding of Multiple Video Programmes*

In this scenario, a common bit budget is divided between different programmes that share the same transmission bandwidth by a joint rate control algorithm. Figure 6.4 shows the diagram block of a typical system ([467]).

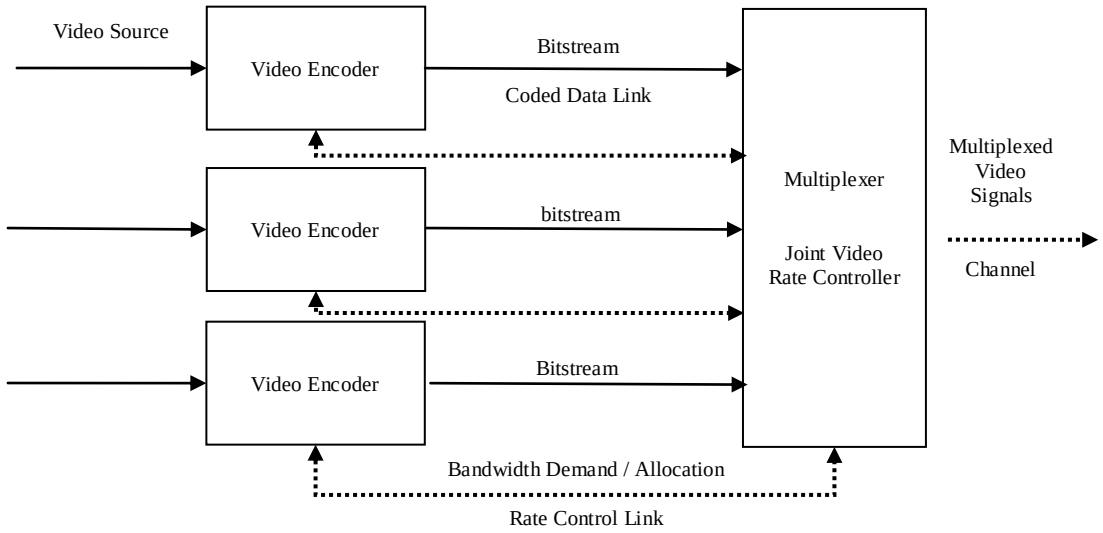


Figure 6.4 – Diagram Block of a Joint Rate Control System

The Multiplexer controls the bit rate of the different video streams (Joint Video Rate Controller). Two connections are set between each Video Encoder and the Multiplexer: the Coded Data Link and the Rate Control Link. The first connection is unidirectional, from the Video Encoder to the Multiplexer, and is used to send the bitstream. The flow of data can vary between tens of kbps up to ten of Mbps depending on the type of programme and the selected encoder (videoconference up to HDTV). The second connection is a bi-directional link of moderately low data-rate. The types of messages are the control type (request bit rate, allocate new bit rate, start adaption to the new bit rate, etc.). Another type of information can be exchanged such as the complexity of encoded bitstreams or the buffer level of the Video Encoder. In a practical implementation, it will be possible to carry rate control information within the bitstream on the Coded Data Link. Thus, the delay between time-critical messages, generated by the Video Encoder, is equal in all video programmes. After receiving information

from the different Video Encoders, the Multiplexer needs to decide how to allocate bandwidth between the Video Encoders. A typical criterion is the complexity of the encoded programmes. A short dialogue takes place over the Rate Control Link.

The Video Encoder needs to adapt the buffer size before change to a new value of the bit rate. The way the buffer is controlled differs when working in CBR or VBR mode ([51],[52]). For CBR coding, changing the number of bits allocated to each picture requires the existence of a buffer in the decoder to store the extra bits. The degree to which an encoder can vary the amount of bits allocated to each picture, depends on the size of this buffer. If the buffer is large, the encoder can use greater variations and thus improve the picture quality at the cost of increasing the decoding delay. The delay is the time taken to fill the input buffer from empty to its current level. A Video Encoder needs to know the size of the decoder's input buffer in order to determine to what extent it can vary the distribution of coding bits among the pictures in a programme. In the case of a constant data transfer rate between a CBR encoder and a decoder, there is a complementary relationship between coder and decoder buffer occupancy: if the encoder buffer is β% full at time δ, the decoder buffer will be (100 - β)% filled at time *(t* + δ).

$$B_e(t) + B_d(t+T) = B \tag{6.1}$$

where the occupancy of the coder buffer at time t is $B_e(t)$, and the occupancy of the decoder buffer at a time t plus δ is $B_d(t+\delta)$. This complementary relationship is very important. It means that if the encoder buffer is prevented from over or under-flowing, then the decoder buffer is guaranteed never to over or underflow. This fact is frequently used to synchronise the decoder. For the decoder and encoder to remain synchronised, the delay through the encoder and decoder, δ, also known as the "codec delay," must remain constant ($\delta = B/r$ where B is the coder and decoder buffer size and r is the data transfer between coder and decoder buffers ([311],[312],[313]). The complimentary relationship is not valid if the data transfer rate between coder and decoder is altered. Thus in order for the codec delay to maintain valid the following equation must hold

$$\delta = \frac{B_c}{r_{\max}} = \frac{B_d}{r_{\min}} \Rightarrow B_c = B_d \frac{r_{\max}}{r_{\min}} \tag{6.2}$$

where B_c and B_d correspond to the size of the coder and decoder buffer (in bits), $r_{\max}$ and $r_{\min}$ are the maximum and minimum value of the bit rate at which the system can function. When working at constant bit rate, r, the upper and lower limit for encoder buffer occupancy, respectively $B_{\max}(r)$ and $B_{\min}(r)$, must be fulfilling, and can be expressed as:

$$B_{\max}(r) = B_d \frac{r}{r_{\min}}$$
$$B_{\min}(r) = B_{\max}(r) - B_d \tag{6.3}$$

During an interval of time immediately before a change in bit rate, corresponding to the codec delay, equation (6.3) is not valid. For the duration of this period, δ, new limits need to be computed by the following expression:

$$B_{\max}(t) = \frac{(t+\delta) B_{\max}(r_{new}) - tB_{\max}(r_{prev})}{\delta}$$
$$B_{\min}(t) = \frac{(t+\delta) B_{\min}(r_{new}) - tB_{\min}(r_{prev})}{\delta} \tag{6.4}$$
$$where - \delta \le t \le 0$$

where t is the time, r_{new} and r_{prev} correspond to the bit rate after and before the period of adaptation start, and $B_{\max}(r_{new}), B_{\max}(r_{prev}), B_{\min}(r_{new})$ and $B_{\min}(r_{prev})$ are computed using Equations (6.3). The Video Encoder needs to adjust the encoder buffer occupancy to new limits and then alters the bit rate. The initiative for changing the bit rate may be forced by the Multiplexer or a request from one of the Videos Encoders. In the second case, the Video Encoder forwards a message requesting the alteration of the bit rate to a new value. This request needs to be authorized and depends on the overall system. If approved, a message is sent to the Video Encoder to start the adjustment of the occupancy limits of the video coder buffer (Equation (6.4)). After receiving the previous message, the Multiplexer waits a set time δ before applying the new bit rate. Figure 6.5 illustrates an example of bit rate adaptation, from 1500 to 2000 kbps, for the coastguard video sequence, in a MPEG 2 Video Encoder ([54]).

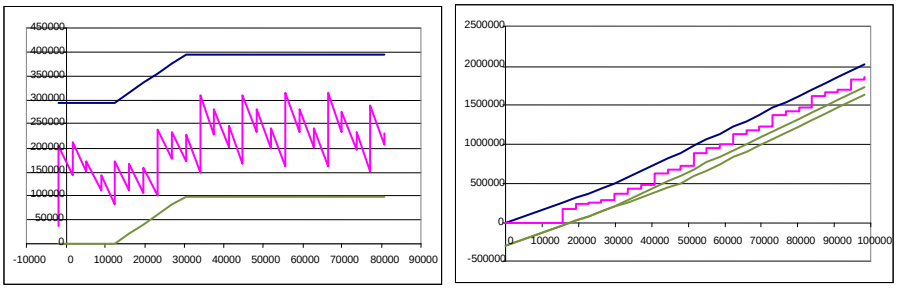


Figure 6.5 – Coder buffer occupancy (left); Address decoder buffer evolution (right)

The Multiplexer and the Video Encoder have a "master-slave" relationship. The Multiplexer monitors all the different Video Encoders and must decide the best way to allocate bandwidth

considering several aspects, including channel capacity or bit rate limits. A similar analysis can be performed when working in VBR mode (Figure 6.6) ([51]).

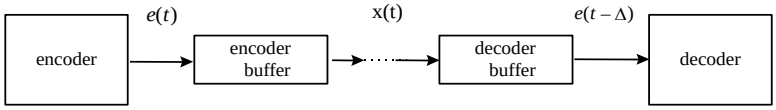


Figure 6.6 – Encoder and Decoder Buffer Diagram

The encoder places bits in the encoder buffer, at the non-uniform rate e(t). Bits are then transferred between encoder and decoder buffers at the rate x(t). If we are encoding in CBR mode, then x(t) is a constant, but in the case of statistical multiplexing x(t) is a time function. Let *E* be the encoder buffer occupancy, *D* the decoder buffer occupancy and B_d the decoder buffer. The encoder buffer occupancy (fullness) is described by the following pair of equations:

$$E(0) = 0 \text{ and } E(t) = \int_0^t (e(\tau) - x(\tau))d\tau \tag{6.5}$$

The decoder waits for an amount of time, before removing the first bit from its buffer. This time is the buffer delay, T (in seconds). The information is retrieved from the decoder buffer in a FIFO order. Let us now consider the decoder buffer occupancy:

$$D(0) = 0 \text{ and } D(t) = \int_0^t x(\tau)d\tau \quad if \quad t < T \tag{6.6}$$

$$D(t) = \int_0^t x(\tau)d\tau - \int_0^{(t-T)} e(\tau)d\tau \quad if \quad t \geq T \tag{6.7}$$

$$D(t+T) = \int_0^{(t+T)} x(\tau)d\tau - \int_0^t e(\tau)d\tau \quad if \quad t \geq T \tag{6.8}$$

By adding the expressions for $E(t)$ and $D(t+T)$ we obtain, for $t \geq T$, the following equation:

$$E(t) + D(t+T) = \int_t^{(t+T)} x(\tau)d\tau \tag{6.9}$$

In order to prevent decoder buffer overflows and underflows, the rate control algorithm must guarantee that

$$0 \leq D(t+T) \leq B_d \tag{6.10}$$

Using equation (6.9) in expression (6.10) we obtain:

$$\begin{cases} 0 \le \int_{t}^{t+T} x(\tau)d\tau - E(t) \\ E(t) \le \int_{t}^{t+T} x(\tau)d\tau \end{cases} \tag{6.11}$$

$$\begin{cases} \int_{t}^{t+T} x(\tau)d\tau - E(t) \le B_d \\ E(t) \ge \int_{t}^{t+T} x(\tau)d\tau - B_d \end{cases} \tag{6.12}$$

Conditions (6.11) and (6.12) give the constraints on the encoder buffer fullness to avoid decoder buffer overflows and underflows in the decoder.

Several joint video encoding methods have been presented in the literature. A brief review is presented here. L. Wang et al. propose a dynamic bit allocation strategy for joint coding that dynamically allocates the available coding bits among the programmes according to the programme related complexities ([35],[36],[37],[38],[56]). In this case, complexity has been defined as the product of a quantisation parameter and the number of bits generated for the frame using the respective quantisation parameter as specified in MPEG-2 TM-5. Mahesh Balakrishnan et al. in [482] discuss the complexity involved in implementing statistical multiplexing proposing a model that defines picture quality. Few results are presented. Boroczky et al. propose an algorithm that dynamically allocates the channel bandwidth among the MPEG-2 programme encoders according to the relative complexities of the programmes using picture and coding statistics ([33],[39]). Jun Xin et al. in [40] presents, for joint transcoding systems of MPEG-1 video sources, a joint bit-allocation approach where the target number of bits for each GOP is proportional to its square-root complexity (complexity is determined following the TM-5 method). Li et al. ([41]) have developed an adaptive joint rate control scheme in which the total bandwidth is, firstly, assigned to each programme according to its complexity so that equal picture quality of the video programmes is maintained. In a second step, for each programme, the corresponding quantisation parameters are then computed by using an adaptive rate control algorithm. The complexity of the programme is determined by its size (number of macroblocks), motion (the sum of motion vector components), and a variable–like measure, like the square of total mean absolute difference (MAD). Simulations have taken place with the MPEG-2 video programme and rate control follows the method presented in [32]. Vincent et al. present a model to calculate the coding gain of resulting joint coding by comparing the bit rates required to achieve a given probability of low picture quality ([42]). The model uses empirical data from MPEG-2 simulations to derive relationships. The

statistical distribution of programme complexity was modelled as a Gamma distribution. Jordan et al. presented methods for the modelling and performance prediction of MPEG-2 video statistical multiplexing of programmes ([43]). Bit rate Cumulative Distribution Function (CDF) data from each channel type were characterised and modelled, the higher part of the CDF by an exponential function, whereas the lower part is better described by a polynomial. The PDF can then be calculated by differentiation. Gu and Lin in [483] outline an approach to R-D optimal solutions, for joint rate control for multiple H.263 video sources, under several common types of distortion measures. These approaches are mainly based on the MPEG-2 coding platform using the same complexity measure as that specified in TM5. However, the TM5 bit rate algorithm was conceived for the MPEG-2, and not H.264/MPEG-4. Work in this area applied to codec's such as H.264/AVC has been the target in the last few years. J. Yang et al. propose in [45],[46] an approach where the mean absolute difference (MAD) of the residual components is used as the complexity measure to adapt to the characteristics of H.264/AVC video coding. Joint allocation concept has been extended to support priorities between different sources being multiplexed. This could be important to video-on-demand servicing. Soon-kak Kwon's has proposed a joint bit rate allocation by using model parameters for MPEG-2 coding. Soon-kak Kwon proposal can be extended to H.263, H.264/AVC and MPEG-4 ([484]).

Tiwari in [485],[486] proposed and compared various methods for allocating bit rate for multiple video streams using dual-frame video coding. Motion activity is used to select if a frame should be encoded as a Long Term Reference (LTR) frame. As no information from future frames is used, this method will fail at a scene cut. A simple RD model (D=a+b/R) is used (MSE is used to measure distortion). Simulations were performed using the H.264/AVC reference software JM 10.1 (baseline profile). The video sequences used for the simulations were are QCIF, 30 fps and of length 300 frames. The first frame is an I-frame and the remaining were coded as P frames. Two reference frames were used. Results for multiplexing two programmes (Akiyo and Foreman) at 60kbps, and four video programmes (Akiyo, Carphone, Coastguard, Grandma) at a combined bit rate of 120 kbps were presented. Proposed methods improve average PSNR quality. Tiwari et al. in [487],[488] described an approach based on a competitive equilibrium bit rate allocation scheme to improve the quality of all the video streams by finding trades between programmes across time. A central rate controller gathers rate-distortion information, at every slot, from each programme and allocates bit rate using an Edgeworth box solution. All computational calculus is performed in the central rate controller. The final bit rate allocation is a Pareto optimal solution. The key aspect of the algorithm is how to estimate the future RD information for a programme. MSE is used to measure distortion, and can be estimate using the following RD curve: D = a + b/(Rate + c) where a,b and c are curve-

fitting coefficients. Simulations were performed using the baseline profile of H.264/AVC JM reference software version 11.0. The GOP size is 15 frames without B-slices. Twelve video SIF sequences, at 30 fps, with duration of 250 seconds were used. The test video sequences were obtained from travel documentaries. For assessing picture quality it was selected the PSNR picture quality metric. Reference bit rate for each programme varied between 190–290 kbps. Results showed an increment in picture quality but the algorithm does not minimise the average distortion. Tiwari et al in [489],[490] proposed a decentralised process based on a pricing mechanism that does not require a heavy computational burden on a central controller. Each user independently is responsible for computing his bit rate demand for the current slot based on current price, available money, and relative video complexity for the current slot compared to the estimated average complexity for future slots. The rate allocator receives the request, normalises the total demand and sends the bit rate price for the next slot based on the total demand and total available bit rate. Various methods of price-based decentralised bit rate allocation are discuused. Results are reported to provide an improve in picture quality for all users. Although centralized allocation slightly outperforms the decentralized allocation, this approac has the advantage of reducing the amount of information shared by the users, and removing the computational burden imposed on the allocator in the centralized approach. The computational complexity grows exponentially with the number of users in the centralized allocation.

Nesrine Changuel ([491],[492],[493],[494],[495],[496]) proposed a statistical multiplexer where a closed-loop control of both video encoders and buffers is performed jointly using a PID feedback. The key idea is to update the encoding rate for each video unit according to the average level of the buffers, to maximize the quality of each programme and effectively use the available channel rate. PSNR was selected to control the video quality. Two RD models were used: a linear RD model for PSNR as a function of QP and an exponential model for rate as a function of QP. To evaluate the performance of the proposed joint encoder and buffer controller, four CIF programmes, 30 fps, were encoded with an H.264/AVC encoder in baseline profile. Intra refresh frame was set to 15, and no B-slices were used (no information regarding number of reference frames, entropy encoding, or RDO is available). To use predictive control, it was selected a window of 1 second. Two scenarios were analysed: a constant channel rate (1Mbps) and a time-varying channel rate. Results show a decrease the intra-programme quality variations compared to the non predictive control.

Zhihai He et al. ([497]) proposed a linear rate model and a linear rate control scheme for H.264/AVC video coding. Based on the linear relationship between the overall bit rate encoding and the fraction of zeros among quantized transform coefficients, it was proposed a linear rate

control for video encoders. The bit rate of the different programmes was controled by a linear scheme using a look-ahead scheme to collect the RD statistics of future frames for joint rate allocation. MSE was selected to measure distortion of the different programmes. Simulations were performed using H.264/AVC JM reference software (version 9.5), CABAC for entropy coding, motion search range of ±16 pixels with up to five reference frames, five TV video clips (CIF size, 30 fps). No B frames and a period of Intra frames of 90 frames. Intra frames of different videos were not synchronized. The total network bandwidth is set to be 7.5 Mbps. The start-up delay is 1 second, and the look-ahead window size is 15 frames (0.5 seconds). Results indicate an average quality improvement about 2-3 dB (PSNR). As video sequences are not available, it is no possible to replicate the experience.

Valenzise et al ([458],[498],[499],[500],[501]) proposed an algorithm to achieve the same distortion for video sequences, using the ρ-domain model. Bandwidth is allocated according to different criteria: minimising the average object distortion (MINAVE) or minimising the variance of the object distortions (MINVAR) under some rate constraint. In both cases, rate-distortion characteristics are described by an exponential model using MSE to measure distortion. To obtain model parameters, input sequences were encoded at a fixed QP=20. Four tests CIF video sequences were used: Foreman, Hall monitor, Soccer and Coastguard. Each sequence contains 300 frames and were encoded at 30fps. The first frame is intracoded (I slices), and the remaining frames were all interframe coded (P slices). No information regarding motion search range, number of reference frames, entropy coding or RDO is available. In the different papers, the bit budget for each sequence is varied between 1.2 bpp, 0.5 bpp and 1/3 bpp (channel bandwidth equals to bpp x number of sequences). Results have shown that, on average, the coding efficiency loss incurred by MINVAR allocation compared with MINAVE is on the order of a fraction of dB, using a PSNR distortion metrics. A similar scheme was presented in [502],[503] for the case of multiple H.264/AVC video sequences (MVS): the frames of the different sequences at a given time instant are grouped in a “multiframe” (MFRM), and a group of multiframes constitutes a “multi-GOP” (MGOP). The coding complexity of each frame is computed using a rho-domain model. No look-ahead window is used. Bits are allocated in such a way that the quality fluctuations between adjacent MGOPs and among the frames inside the MGOPs are minimized. The performance of MVS was analysed using three CIF test sequences at 30 Hz (Foreman, News, Mobile) and with 300 frames. The GOP size was set to 15 and the GOP pattern was IBBPBB... No information regarding motion search range, number of reference frames, entropy coding or RDO is available. Four values for the channel bandwidth were selected: 750 kbps, 1500 kbps, 3000 kbps and 6000 kbps. Results show a major reduction in PSNR variation.

In summary, a wide array of joint video coding / statistical multiplexing schemes have been developed. The results have shown this is an efficient method to obtain uniform picture quality among video programmes, while maintaining the aggregate bit rate of the various video programmes conforming to the channel capacity. Nevertheless, the algorithms use mainly PSNR and MSE picture quality metrics to assess picture quality. As mention in Chapter 2, this type of metrics fails to predict the HVS perception because they take no account of where errors occur in the image, not every change in an image is noticeable or leads to distortion, no error is visually important. In the present work, two perceptual video quality metrics were selected, the Structural SIMilarity (SSIM) index and the JND (Just Noticeable Distortion). A novel approach to a joint video source coding, based on perceptual distortion, is thus proposed. The final aim of an encoding system is to ensure that as many viewers as possible are satisfied with the quality of the programme, they have selected. In the present joint video encoding studies, no subjective video quality assessment of the results has been performed yet.

6.2 Methods for Joint Video Encoding

Among existing solutions for implementing joint video coding, one of the most popular allocation methods is based on the complexity defined in the rate control of MPEG-2 TM5 ([48],[49],[50],[51],[52],[53],[54],[504],[505]). Most of the existing solutions that can be found in the literature are oriented to joint coding of MPEG-2 video streams, in the majority of cases, using a feed-forward strategy (Figure 6.7). In a first stage, video pictures of the different video programmes are encoded using a fixed quantisation parameter. Statistics from the first stage, such as the bit rate and distortion, are extracted.

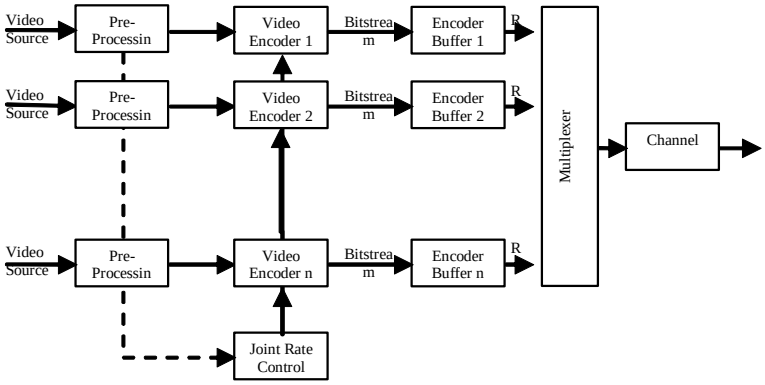


Figure 6.7 – Block Diagram for Joint Coding of Video Programmes

In the second stage, these statistics are used to estimate the video programme's complexity and improve the bit rate control process. Coding parameters, such as spatial and temporal resolution, GOP pattern and motion estimation techniques, are kept constant in both the first and the second

stage. Usually, the value of the quantisation parameter in the first stage varies between 15 and 25.

In the literature, this approach has been extended into some of the current H.264/AVC proposals, based on picture quality metrics. The MPEG-2 joint video coding and statistical multiplexing algorithm is composed of four steps. In the first step, the reference bandwidth BW_{ref} of each video programme is calculated. It consists of the bandwidth that would be necessary to encode a frame of a GOP when each video programme is encoding at CBR and independently of the other's video programmes. The reference bandwidth is determined based on the total available transmission bandwidth, the picture coding complexity and picture type, GOP pattern and the current state of the total virtual buffer. It can be determined as follows, depending on the picture type:

$$BW_{ref\_I,p} = \frac{X_{i,p}}{X_{i,p} + \frac{N_{p,p}X_{p,p}}{K_{p,p}} + \frac{N_{b,p}X_{i,p}}{K_{b,p}}} \times R_p \tag{6.13}$$

$$BW_{ref\_P,p} = \frac{\frac{X_{p,p}}{K_{p,p}}}{X_{i,p} + \frac{N_{p,p}X_{p,p}}{K_{p,p}} + \frac{N_{b,p}X_{i,p}}{K_{b,p}}} \times R_p \tag{6.14}$$

$$BW_{ref\_B,p} = \frac{\frac{X_{b,p}}{K_{b,p}}}{X_{i,p} + \frac{N_{p,p}X_{p,p}}{K_{p,p}} + \frac{N_{b,p}X_{i,p}}{K_{b,p}}} \times R_p \tag{6.15}$$

where $BW_{ref\_I,p}$, $BW_{ref\_P,p}$ and $BW_{ref\_B,p}$ are the bandwidth references for I, P and B pictures for the pth video programme, $K_{p,p} = 1.0$ and $K_{b,p} = 1.4$ are constants for P and B pictures of the pth video programme that depend on the quantisation matrices; $N_{p,p}$ and $N_{b,p}$ are the number of P and B pictures remaining for pth video programme in the current GOP in encoding order and R_p is the transmission bandwidth for the pth video programme allocated to the channel during one GOP being computed as in Equation (6.16)

$$R_p = \frac{N_{GOP,p} \times bit\_rate_p}{f_p} \tag{6.16}$$

where $N_{GOP,p}$, f_p and bit_rate_p are respectively the total numbers of pictures in the current GOP, the frame rate and the bit rate for the pth video source. X variables refer to the overall complexity of the different picture types and are updated by calculating the product of the number of bits generated by encoding a picture and the average quantisation parameter (computed with the actual quantisation values used during the encoding of all macroblocks) for each of the various picture types (Equation (6.17)).

$$X' = R' \times Q' \tag{6.17}$$

Initial values of X variables are computed according to the next equation:

$$X_i = \frac{160 \times bit\_rate}{115}, X_p = \frac{60 \times bit\_rate}{115}, X_b = \frac{42 \times bit\_rate}{115}, \tag{6.18}$$

The goal in the second step is to obtain a measure of the complexity of the different video programmes. This can be achieved by coding a complete GOP or just a simple frame of each video programme using a fixed quantisation parameter. Using Equation (6.17) the estimated bandwidth, BW_{Est_p}, for each p[th] video programme can be determined as

$$\begin{aligned} X_p &= BW_{Est_p} \times Q_p \\ BW_{Est_p} &= \frac{X_p}{Q_p} = \frac{X_p}{Q_{Q=Fixed}} \end{aligned} \tag{6.19}$$

In the third step, the reference bandwidth is allocated to each video programme proportionally to the value estimate in the second step

$$BW_p = \frac{BW_{Est_p}}{\sum_{p=1}^{M} BW_{Est_p}} \times \sum_{p=1}^{M} BW_{ref_p} = \frac{\frac{X_p}{Q_p}}{\sum_{p=1}^{M} \frac{X_p}{Q_p}} \times \sum_{p=1}^{M} BW_{ref_p} = \frac{X_p}{\sum_{p=1}^{M} X_p} \times \sum_{p=1}^{M} BW_{ref_p} \tag{6.20}$$

where *M* is the number of video sources. As the quantisation parameter is equal for all video programmes, the final equation can be simplified, and bandwidth is allocated proportionally to programme complexity. Finally, in the fourth and last step, a reference value for the quantisation parameter is determined for each picture. The reference value of the quantisation parameter is then independently modulated in each video encoder, according to the spatial activity in the macroblock, to obtain the quantisation parameter, mquant, used to quantise each macroblock ([19]).

6.2.1 *Joint Video Encoding and TM5 Complexity Metric (Mux Bit)*

A first approach to the joint coding of H.264/AVC video programmes is to extend the algorithm based on the TM5 definition of complexity (Equation (6.17)). This algorithm uses a feed forward rate control and is composed of four steps:

- First-pass coding using a fixed quantisation parameter.
- Calculation of reference bandwidth.
- Reallocate bandwidth per video programme according to its needs.
- Encode the frame and update statistics.

Before performing the first-pass, it is necessary to select the common quantisation parameter value. Two methods are possible: to select a unique value for all the sequence or to select an initial value and define an update method. The first method is usually used when the GOP pattern is not known in advance or when video programmes do not have the same size or are not GOP aligned. The possibility of updating the fixed quantisation parameter allows better statistics to be obtained. In the present method, an update method is defined based on the assumption that GOP size is equal in all video programmes, and that GOP are aligned. Quantisation parameters are defined for each picture type, based on previous quantisation parameters values and GOP pattern.

For GOP patterns without B frames, the quantisation parameter is determined as the average value of the previous quantisation parameter's values of the preceding GOP. Otherwise, for GOPs containing B frames, the average quantisation parameter from the previous GOP, from all video programmes, is first determined and then subtract one for the I and P frames. For B frames, instead of subtracting the value of one is added.

In the second step, the reference bandwidth is estimated as the value that would be allocated in case of independent coding. Let $N_{gop,p}$ designate the total number of frames in a GOP for the pth video programmes, $n_{i,j,p}(i=1,2,\cdots, j=1,2,\cdots, N_{gop,p}, p=1,2,\cdots M)$ refer to the jth frame in the ith GOP of the pth video programme, and $B_c(n_{i,j,p})$ represent the occupancy of virtual buffer of the pth video programme after encoding the jth frame in the ith GOP. To determine the value of the virtual occupancy of the buffer the following expression can be used

$$\begin{aligned} B_c(n_{i,j+1,p}) &= B_c(n_{i,j,p}) + b(n_{i,j,p}) - \frac{\mathrm{R}(n_{i,j,p})}{f} \\ B_c(n_{1,1,p}) &= 0 \\ B_c(n_{i+1,0,p}) &= B_c(n_{i,N_{gop},p}) \end{aligned} \tag{6.21}$$

where $b(n_{i,j,p})$ is the number of bits generated by the j[th] frame in the i[th] GOP of the p[th] video programme, $\mathrm{R}(n_{i,j,p})$ is the available bit rate at the j[th] frame in the i[th] GOP of the p[th] video programme, and f is the frame rate (it is assumed an equal value for all video programmes). In the beginning of the i[th] GOP of the p[th] video programme, the total number of bits allocated for the GOP is determined by

$$T_r(n_{i,0,p}) = \frac{R(n_{i,1,p})}{f} \times N_{gop} - B_c(n_{i-1,N_{gop},p}) \tag{6.22}$$

T_r is updated on a frame basis according to Equation (6.23)

$$T_r\left(n_{i,j,p}\right) = T_r\left(n_{i,j-1,p}\right) + \frac{R\left(n_{i,j,p}\right) - R\left(n_{i,j-1,p}\right)}{f} \times \left(N_{gop} - j\right) - b\left(n_{i,j-1,p}\right) \tag{6.23}$$

Therefore, the aggregate bandwidth of all the video programmes when starting encoding the j[th] frame of the i[th] GOP can be determined as follows:

$$T_{aggregate_{i,j}} = \sum_{p=1}^{M} T_r\left(n_{i,j,p}\right) \tag{6.24}$$

Equation (6.24) is used as reference bandwidth. It corresponds to the total number of remaining bits when starting encoding the j[th] frame of the i[th] GOP of all video programmes.

The next step is to calculate the new value of the bandwidth for each video programme. The criterion for allocating bandwidth is to be proportional to the ratio complexity between programme complexity and overall video programmes' complexity. Thus, the available bandwidth at the beginning of the j[th] frame of the i[th] GOP of the p[th] video programme can be expressed as follows:

$$T_{i,j,p} = \frac{X_{i,j,p|Q=fixed}}{\sum_{p=1}^{M} X_{i,j,p|Q=fixed}} \times T_{aggregate_{i,j}} = \frac{T_{i,j,p|Q=fixed}}{\sum_{p=1}^{M} T_{i,j,p|Q=fixed}} \times T_{aggregate_{i,j}} \tag{6.25}$$

where $X_{i,j,p|Q=fixed}$ and $T_{i,j,p|Q=fixed}$ are, respectively, the complexity and the bit rate obtained when coding using a fixed quantisation parameter, and M the number of video programme that are jointly encoded. This approach is denominated as "Mux Bit" from this point forward. Mux Bit Joint Video Coding can be summarizing in the following steps:

Step 1) Initialization: At the beginning of the video encoding session, define the number of video programmes, channel bandwidth, reference video bit rate per video programme, and encoder parameters such as GOP pattern, motion parameters, buffer size.

Step 2) Look-ahead process. The quantisation parameter is determined according to the GOP pattern. Compute the average value of the quantisation parameter of the previous GOP. If the sequence is encoded with a GOP pattern without B frames, then use this value to encode I and P pictures. Otherwise, if GOP pattern has B frames then subtract value one from the average value of the quantisation parameter, for I and P frames, or if frame type is B add value one to the average value of the quantisation parameter. For the first GOP, use Equation (4.63).

Step 3) Encode a frame of each video programme using the fixed quantisation parameter. Store statistics.

Step 4) Determine the reference bandwidth per video programme (Equation (6.24)).

Step 5) Determine the joint coding bandwidth per video programme (Equation (6.25)). Check if the allocate bandwidth generate a potential buffer overflow/underflow.

Step 5) Encode one frame of each video programme used according to picture level rate control as explained in section 4.2.4.2.

Step 6) Loop until encode GOP: Increase the frame number by one and go to step 3.

Step 7) Loop until encode the sequence: Increase the GOP number by one and go to 2.

6.2.2 *Joint Video Encoding with R-D Models*

In the previous Chapter, several R-D models were analysed using traditional objective picture quality metrics and perceptual quality metrics. It was found that regarding PSNR, PSPNR and SSIM, the quadratic function presented good results in modelling R-D. In this section, a new approach will be presented for joint video coding based on these results. H.264/AVC JM encoders use a quadratic rate-distortion model to calculate the corresponding quantisation parameter, which is then used for rate-distortion optimization for each macroblock in the current basic unit. Analysing how the microscopic control inside a codec is performed, it can be noted that, within a frame, the relationship between the quantisation parameter, distortion and the bit rate of a macroblock i is determined by the following rule:

$$R_i = \left(\frac{c_1}{Q_i} + \frac{c_2}{Q_i^2} \right) \times D_i \Leftrightarrow \frac{R_i}{D_i} = \frac{c_1}{Q_i} + \frac{c_2}{Q_i^2} \tag{6.26}$$

In relation to H.264/AVC JM, the distortion metric used is MAD and in the case of MPEG-4 VM8, the distortion is SAD. When encoding a frame, the sum of the bits generated by all the macroblocks should not exceed the number of allocated bits of the frame. The problem regarding rate control is how to distribute the available bits per each macroblock. If it is considered that all macroblocks in a frame are equally important, then each macroblock should be quantised with the same quantisation parameter ($Q_1 = \cdots = Q_n$). As a result,

$$\begin{cases} \dfrac{R_i}{D_i} = \dfrac{c_1}{Q_i} + \dfrac{c_2}{Q_i^2} \Rightarrow \dfrac{R_1}{D_1} = \cdots = \dfrac{R_i}{D_i} \\ Q_1 = \cdots = Q_n \end{cases} \tag{6.27}$$

$$\sum_i R_i = T_{frame} \tag{6.28}$$

where R_i, D_i, Q_i are respectively the number of bits, the distortion and the quantisation parameter of macroblock i, c_1 and c_2 are coefficient parameters and T_{frame} is the number of bits available to encode the frame. Solving equation (6.28) by using equation (6.27) it follows that

$$\begin{aligned} R_1 + R_2 .. + R_i &= D_1 \times \frac{R_i}{D_i} + D_2 \times \frac{R_i}{D_i} .. + R_i = \frac{\left(D_1 + D_2 + .. + D_i\right)}{D_i} \times R_i = T_{frame} \\ R_i &= \frac{D_i}{\sum_{k=1}^{n} D_k} T_{frame} \end{aligned} \tag{6.29}$$

Thus, bits should be allocated, on a macroblock basis, proportionally to the impact of the macroblock distortion regarding the overall distortion resulting from all the macroblocks. When performing joint rate video allocation, all video programmes are equally important. In the case of the PSNR, PSPNR and SSIM picture quality metrics, rate distortion can be modelled by the quadratic function with good results. Therefore, in the first pass, using a fixed quantisation parameter it follows that (Equation (6.30)):

$$Q_1 = \cdots = Q_M = Q_{fixed} \Rightarrow \frac{R_{p|Q=fixed}}{D_{p|Q=fixed}} = \frac{c_{p,1}}{Q_{fixed}} + \frac{c_{p,2}}{Q_{fixed}^2} \tag{6.30}$$

where $R_{p|Q=fixed}$ and $D_{p|Q=fixed}$ are, respectively, the bit rate and the distortion of the pth video programme using the fixed quantisation parameter, and $c_{p,1}$ and $c_{p,2}$ are two coefficients for p^{th} programme that depend on the nature of the video programme. When the goal is to obtain uniform distortion among video programmes, it follows:

$$D = D_1 = \cdots = D_p$$
$$\frac{R_p}{D} = \frac{c_{p,1}}{Q_p} + \frac{c_{p,2}}{{Q_p}^2} \tag{6.31}$$

By combining Equation (6.30) and Equation (6.31), Equation (6.32) is obtained

$$\left.\begin{array}{c} \dfrac{R_p}{D} = \dfrac{c_{p,1}}{Q_p} + \dfrac{c_{p,2}}{{Q_p}^2} \\ D = D_1 = \cdots = D_p \\ \dfrac{R_{p,Q=fixed}}{D_{p,Q=fixed}} = \dfrac{c_{p,1}}{Q_{fixed}} + \dfrac{c_{p,2}}{{Q_{fixed}}^2} \end{array}\right\} \frac{R_p}{D} = \frac{\dfrac{\left(c_{p,1} \times Q_p + c_{p,2}\right)}{{Q_p}^2}}{\dfrac{\left(c_{p,1} \times Q_{fixed} + c_{p,2}\right)}{{Q_{fixed}}^2}} \frac{R_{p,Q=fixed}}{D_{p,Q=fixed}} \tag{6.32}$$

Solving in order of D

$$D = \frac{\dfrac{\left(c_{p,1} \times Q_{fixed} + c_{p,2}\right)}{{Q_{fixed}}^2}}{\dfrac{\left(c_{p,1} \times Q_p + c_{p,2}\right)}{{Q_p}^2}} \frac{D_{p,Q=fixed}}{R_{p,Q=fixed}} R_p = \frac{\left(c_{p,1} \times Q_{fixed} + c_{p,2}\right)}{\left(c_{p,1} \times Q_p + c_{p,2}\right)} \frac{{Q_p}^2}{{Q_{fixed}}^2} \frac{D_{p,Q=fixed}}{R_{p,Q=fixed}} R_p \tag{6.33}$$

In order to simplify equation (6.33) α_p is defined as

$$\alpha_p = \frac{\left(c_{p,1} \times Q_{fixed} + c_{p,2}\right)}{\left(c_{p,1} \times Q_p + c_{p,2}\right)} \frac{{Q_p}^2}{{Q_{fixed}}^2} \frac{D_{p,Q=fixed}}{R_{p,Q=fixed}} \tag{6.34}$$

then distortion can be determined as follows

$$D = \alpha_p \times R_p \tag{6.35}$$

Note that α_p can be further simplified if it is considered that $c_{p,1} \times Q_{fixed} \gg c_{p,2}$

$$\alpha_p = \frac{Q_p}{Q_{fixed}} \frac{D_{p,Q=fixed}}{R_{p,Q=fixed}} \tag{6.36}$$

Equation (6.35) is valid for any image in the video programme so that it can be written as a function of the jth frame of the ith GOP of the pth video programme (equation (6.37)):

$$D_{i,j} = \alpha_{i,j,p} \times R_{i,j,p} \tag{6.37}$$

Considering the goal of making distortion uniform, then

$$D = D_{i,j,1} = .. = D_{i,j,p} \Leftrightarrow \alpha_{i,j,1} R_{i,j,1} = \alpha_{i,j,2} R_{i,j,2} = .. = \alpha_{i,j,p} R_{i,j,p}$$
$$\sum_{p=1}^{M} R_{i,j,p} = R_{i,j,1} + R_{i,j,2} .. + R_{i,j,p} = T_{aggregate_{i,j}} \tag{6.38}$$

The allocation process is constrained by the channel capacity. Thus, the sum of the allocated bandwidth for all video programmes cannot exceed the channel bandwidth

$$\sum_{p=1}^{M} R_{i,j,p} = R_{i,j,1} + R_{i,j,2} .. + R_{i,j,p} = T_{aggregate_{i,j}} \tag{6.39}$$

Combining Equation (6.38) with Equation (6.39) results in

$$\frac{\alpha_{i,j,p}}{\alpha_{i,j,1}} R_{i,j,p} + \frac{\alpha_{i,j,p}}{\alpha_{i,j,2}} R_{i,j,p} + .. + R_{i,j,p} = \alpha_{i,j,p} \left(\frac{1}{\alpha_{i,j,1}} + \frac{1}{\alpha_{i,j,2}} + .. + \frac{1}{\alpha_{i,j,p}} \right) R_{i,j,p} = T_{aggregate_{i,j}} \tag{6.40}$$

$$R_{i,j,1} + R_{i,j,2} .. + R_{i,j,p} = \frac{\alpha_{i,j,p}}{\alpha_{i,j,1}} R_{i,j,p} + \frac{\alpha_{i,j,p}}{\alpha_{i,j,2}} R_{i,j,p} + .. + R_{i,j,p} =$$
$$= \alpha_{i,j,p} \left(\frac{1}{\alpha_{i,j,1}} + \frac{1}{\alpha_{i,j,2}} + .. + \frac{1}{\alpha_{i,j,p}} \right) R_{i,j,p} = T_{aggregate_{i,j}} \tag{6.41}$$

In order to simplify the equation (6.41) let $\beta_{i,j,p}$ be the inverse of $\alpha_{i,j,p}$

$$\beta_{i,j,p} = \frac{1}{\alpha_{i,j,p}} \tag{6.42}$$

Replacing $\alpha_{i,j,p}$ by $\beta_{i,j,p}$ in Equation (6.41) it is obtained

$$R_{i,j,p} = \frac{\beta_{i,j,p}}{\sum_{k=1}^{M} \beta_{i,j,k}} T_{aggregate_{i,j}}$$
$$\beta_{i,j,p} = \frac{1}{\alpha_{i,j,p}} \tag{6.43}$$

To compute Equation (6.43) it is necessary to use data from first pass (rate, distortion and quantisation parameter) and Q_p. As was shown in previous Chapter, it is possible to model the perceptual picture quality metrics such as PSNR, SSIM or PSPNR as a function of the quantisation parameter (D-QP model). As a result, for a specific picture quality value, the quantisation parameter can be obtained using the D-QP function. To estimate the picture quality it must be considered that the variation of the quality between neighbouring frames should be

reduced. Thus, it is proposed that its value be obtained as a weighted average of the picture quality values of the previous GOP frames. From Chapter 5, the best functions to model D-QP are the quadratic and the linear approach. The linear regression process can be represented by:

$$D_i = l_1 \times Q_i + l_2 \tag{6.44}$$

and l_1 nd l_2 are parameters obtained using the least square method by setting its value to zero (E) after partially differentiating with regard to l_1 nd l_2 (Equation (6.45)).

$$\begin{aligned} E &= \sum \left(D_i - l_1 \times Q_i - l_2\right)^2 \\ \frac{\partial E}{\partial l_1} &= \sum \left(D_i - l_1 - l_2 \times Q_i\right) \times 2 \times \left(-Q_i\right) = 0 \\ \frac{\partial E}{\partial l_2} &= \sum \left(D_i - l_1 - l_2 \times Q_i\right) \times 2 \times \left(-1\right) = 0 \end{aligned} \tag{6.45}$$

Thus, the solution is

$$\begin{aligned} l_1 &= \frac{\sum(1)\sum\left(Q_i \times D_i\right) - \sum Q_i \sum D_i}{\sum(1)\sum\left(Q_i^2\right) - \left(\sum Q_i\right)^2} \\ l_2 &= \frac{\sum\left(D_i\right)\sum\left(Q_i^2\right) - \sum Q_i \sum\left(Q_i \times D_i\right)}{\sum(1)\sum\left(Q_i^2\right) - \left(\sum Q_i\right)^2} \end{aligned} \tag{6.46}$$

In the case of the quadratic approach, by applying the linear regression it follows:

$$D_i = l_1 \times Q_i^2 + l_2 \times Q_i + l_3 \tag{6.47}$$

Thus by partially differentiating E according to l_1, l_2, and l_3 respectively and setting to zero:

$$\begin{aligned} \left(\sum Q_i^4\right) l_1 + \left(\sum Q_i^3\right) l_2 + \left(\sum Q_i^2\right) l_3 &= \sum Q_i^2 \times D_i \\ \left(\sum Q_i^3\right) l_1 + \left(\sum Q_i^2\right) l_2 + \left(\sum Q_i\right) l_3 &= \sum Q_i \times D_i \\ \left(\sum Q_i^2\right) l_1 + \left(\sum Q_i\right) l_2 + \left(\sum 1\right) l_3 &= \sum D_i \end{aligned} \tag{6.48}$$

By solving the equation system of (6.48) the following solution is obtained.

$$\begin{aligned} l_1 = {} & \frac{(\sum 1 \sum x^2 \sum x^2 y - \sum x \sum x \sum x^2 y + \sum x \sum x^2 \sum xy}{2\sum x \sum x^2 \sum x^3 + \sum 1 \sum x^2 \sum x^4 - \sum x \sum x \sum x^4 - \sum 1 \sum x^3 \sum x^3 - \sum x^2 \sum x^2 \sum x^2} \\ & + \frac{-\sum 1 \sum x^3 \sum xy + \sum x \sum x^3 \sum y - \sum x^2 \sum x^2 \sum y)}{2\sum x \sum x^2 \sum x^3 + \sum 1 \sum x^2 \sum x^4 - \sum x \sum x \sum x^4 - \sum 1 \sum x^3 \sum x^3 - \sum x^2 \sum x^2 \sum x^2} \end{aligned} \tag{6.49}$$

$$l_2 = \frac{\sum x \sum x^2 \sum x^2 y - \sum 1 \sum x^3 \sum x^2 y + \sum 1 \sum x^4 \sum xy}{2\sum x \sum x^2 \sum x^3 + \sum 1 \sum x^2 \sum x^4 - \sum x \sum x \sum x^4 - \sum 1 \sum x^3 \sum x^3 - \sum x^2 \sum x^2 \sum x^2} + \frac{-\sum x^2 \sum x^2 \sum xy + \sum x^2 \sum x^3 \sum y - \sum x \sum x^4 \sum y}{2\sum x \sum x^2 \sum x^3 + \sum 1 \sum x^2 \sum x^4 - \sum x \sum x \sum x^4 - \sum 1 \sum x^3 \sum x^3 - \sum x^2 \sum x^2 \sum x^2} \quad (6.50)$$

$$l_3 = \frac{-\sum x^2 \sum x^2 \sum x^2 y + \sum x \sum x^3 \sum x^2 y - \sum x \sum x^4 \sum xy}{2\sum x \sum x^2 \sum x^3 + \sum 1 \sum x^2 \sum x^4 - \sum x \sum x \sum x^4 - \sum 1 \sum x^3 \sum x^3 - \sum x^2 \sum x^2 \sum x^2} + \frac{\sum x^2 \sum x^3 \sum xy - \sum x^3 \sum x^3 \sum y + \sum x^2 \sum x^4 \sum y}{2\sum x \sum x^2 \sum x^3 + \sum 1 \sum x^2 \sum x^4 - \sum x \sum x \sum x^4 - \sum 1 \sum x^3 \sum x^3 - \sum x^2 \sum x^2 \sum x^2} \quad (6.51)$$

where

$x = Q_i, y = D_i$

Due to its simplicity and lower complexity compared with the quadratic approach, it was decided to use a linear function to model D-QP. Q_p is used as a reference for the allocation process between video programmes. During the linear regression process, as more and more images are encoded and used to update the model parameters, the sensitivity to new data is progressively reduced. Thus, a threshold value regarding the number of images that should be used in the update process needs to be defined. In addition, at the start of the linear regression process, the update model parameters can be unstable due to the lack of available data. To deal with these conditions, it is proposed that the update process only be started after having finished encoding the first GOP and the data from one GOP used. For each video programme, it is necessary to compute a linear function to model D-QP for each picture type (Equation (6.52))

$$\begin{aligned} Di_{i,j,p} &= li_1 + li_2 \times Qi_{i,j,p} \quad pict\_type = I \\ Dp_{i,j,p} &= lp_1 + lp_2 \times Qp_{i,j,p} \quad pict\_type = P \\ Db_{i,j,p} &= lb_1 + lb_2 \times Qb_{i,j,p} \quad pict\_type = B \end{aligned} \quad (6.52)$$

where $li_1, li_2, lp_1, lp_2, lb_1$ and lb_2 are coefficients of the least square's method. Note that in Equation (6.52), the D-QP functions refer to the ith image of a specific type (I – intra; P – predicted and B - interpolated) of the jth GOP of the pth video programme.

To obtain $Q_{i,j,p}$ a value for the target distortion is needed. PSNR, PSPNR, and SSIM were used as a distortion metric giving origin to three joint video coding algorithms: Mux PSNR based on PSNR, Mux PSPNR based on PSPNR, and Mux SSIM based on the SSIM picture quality metric. These enhanced Joint Rate Video control algorithms can be described in the following steps:

Step 1) Initialization: identify the number of video programmes, channel bandwidth, the reference bit rate per video programme, and encoder parameters such as GOP pattern, motion parameters, buffer size.

Step 2) Look-ahead processing: Encode one GOP of each video programme with fixed quantise parameter. The quantisation parameter is selected as follows.

- If it is the first GOP then use Equation (4.63).
- For the remaining GOPs, compute the average value of the quantisation parameter of the previous GOP. Use this value if the GOP structure has no B frames. Otherwise, for I and P frames, add value one to the average quantisation parameter, and for B frame, subtract value one from the average quantisation parameter.
- Store statistic

Step 3) Compute the aggregate bandwidth of all the video programmes when starting encoding the j^{th} frame of the i^{th} GOP (Equation (6.24)).

Step 4) Compute average picture quality from previous GOP frames of video programmes and use it to estimate a quantisation parameter per video source, Q_p (Equation (6.52)).

Step 5) Compute $\beta_{i,j,p}$ (Equation (6.42)).

Step 6) Allocate bandwidth for each video programme according to Equation (6.43).

Step 7) Encode one frame of each video programme using the rate control at the picture level as explained in section 4.2.4.2.

Step 8) Loop: Update parameters. Increase the frame number by one. Go to step three and repeat steps until all frames in the GOP are encoded.

Step 9) Loop: Increase the GOP number by one. If last GOP stop, otherwise go to step two.

6.3 Objective Video Quality Assessment

In the literature, joint video rate coding algorithms have been assessed using different video sequences, number of video programmes, video resolutions and coding parameters such as bit rate or GOP pattern. Jiang ([45],[46]) evaluates its proposal of a joint rate allocation algorithm for H.264/AVC by using two groups of video sequences. The first group contains four CIF sequences, and the second group contains six QCIF sequences. All the sequences are well-known and available on the web ([436]). For each group, the sequences were encoded

independently and jointly at 15Hz with no B frames. The total channel bandwidth for group one was 1024 kbps (256 kbps per sequence) and 288kbps (48 kbps per sequence) for group two. Information regarding GOP size is not reported. Zhihai ([497]) proposes an H.264/AVC joint rate allocation algorithm. Zhihai simulations used five television clips with CIF resolution, encoded at 30 fps, with a GOP size of 90 frames and a total network bandwidth of 7.5 Mbps. Both Jiang and Zhihai use, in their simulations, CIF sequences but use different values for the channel bandwidth (Jiang 256kbps and Zhihai 1.5Mbps per CIF video programme). Zhihai also proposed a linear rate control for the H.264/AVC coder to be used with the joint rate allocation method ([497]). To evaluate the performance of the new linear rate control, Zhihai used GOP patterns without B frames in the simulations. Section 6.1 includes an analysis to current simulation's scenarios that can be found in the literature. This variation in encoding parameters also occurred in MPEG-2. Soon-kak Lwon ([484]) and Lilla ([39]), evaluate their algorithms using four video programmes with CCIR 601 spatial resolution and varying the channel bandwidth from 16 up to 32 Mbps. Soon-kak Lwon proposes to encode sequences using a GOP structure of 15 pictures with no B-pictures. Lilla has selected for GOP length values of 13 and 16 pictures with two B pictures located between anchor pictures ([39]).

To evaluate the performance of the proposed joint rate allocation methods, Mux Bit, Mux PSPNR, Mux SSIM and Mux PSPNR, several experiments were carried out via simulation. Four scenarios were identified: independent coding and three joint rate-coding algorithms that combine two video sources, three video sources and six video sources. For each scenario, different channel bandwidth was defined (512 kbps and 1024 kbps for two video programmes; 768 kbps and 1536 kbps for three video programmes; and 1536 and 3072 kbps for six video programmes). These scenarios correspond to a bit rate reference per video programme of 256kbps and 512kbps (from this point forward the reference bit rates will be used to distinguish the two-channel bandwidth scenarios).

Cernak et al. in [506] addressed the relationship between video quality, screen resolution, and bit rate. Using the data obtained during two VQEG projects (MM Test and HD Test Phase I) they have plotted, for different screen resolutions (QCIF, CIF, VGA, and HD), MOS as a function of bit rate. Preliminary results show that relation is regular, suggesting that interpolation across screen resolution might be reasonable.

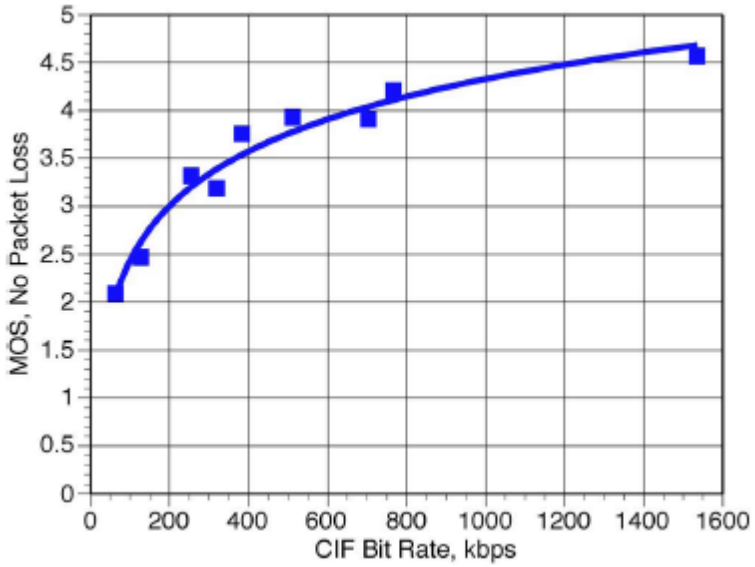


Figure 6.8 – CIF results from VQEG MM project (H.264, no packet loss) ([506])

Looking at Figure 6.8, sequence's results with MOS values between Fair and Good correspond to bit rates between 200kbps and roughly 600 kbps. Although results are based on the largest and broadest video quality testing (for each screen resolution, a set of 30 video sequences was evaluated across all 13 or 14 labs participating in the test), some limitations exist in its analysis. The assessment process was conducted in lab viewing conditions so the VQEG data should represent viewers' judgments at their most critical. Thus, the bit rate value for a given MOS may, in fact, represent an upper bound.

Before selecting the video sequences for the different scenarios, each video sequence was encoded individually using H.264/AVC JM 11.0 baseline-profile encoder ([169],[170]), at two different constant bit rates (256 kbps and 512 kbps) and using 4 different GOP patterns (IPPP GOP1, IPPP GOP2, IBBP GOP1, IBBP GOP2) (section 6.3.1). Sequences were encoded according to the parameters defined in Table 5.2. After analysing the performance of video programmes according to picture quality (PSNR, PSPNR and SSIM) and CoV, it was selected three sequences for the first two multiplexing scenarios (two and three sources) (section 6.3.2) and a group of six video sources for the third multiplexing scenario (section 6.3.3).

6.3.1 *Independent Video Encoding Performance Analysis*

Three different picture quality metrics were used to assess results: PSNR, PSPNR and SSIM. For each picture quality metric, the mean, the standard deviation and Coefficient of Variation (CoV) of the different picture quality metrics for each video programme are presented. CoV, also known as "relative variability," is the ratio of the standard deviation to the mean. It is a

helpful descriptive statistic since the standard deviation of data can only be properly understood in the context of the mean of the data.

	IPPP_GOP1			IPPP_GOP2			IBBP_GOP1			IBBP_GOP2		
Video	Mean	Std	CoV	Mean	Std	CoV	Mean	Std	CoV	Mean	Std	CoV
aki	37.9	0.58	0.02	41.2	1.09	0.03	42.6	1.87	0.04	41.7	1.55	0.04
cgd	27.9	0.76	0.03	29.0	0.94	0.03	30.3	0.96	0.03	29.9	0.96	0.03
dea	28.9	0.86	0.03	33.0	1.00	0.03	37.2	1.57	0.04	35.0	1.22	0.03
flg	22.2	1.58	0.07	25.0	1.12	0.04	27.3	1.17	0.04	26.1	1.14	0.04
for	31.1	2.21	0.07	33.1	1.84	0.06	33.9	1.62	0.05	33.3	1.67	0.05
fot	28.5	3.86	0.14	29.2	3.57	0.12	29.2	3.67	0.13	29.1	3.48	0.12
hal	33.2	0.50	0.01	36.2	0.44	0.01	38.2	0.88	0.02	37.5	0.79	0.02
mad	34.7	0.70	0.02	37.3	1.50	0.04	39.0	2.26	0.06	38.1	1.63	0.04
mcl	20.6	1.55	0.08	23.6	1.05	0.04	26.8	0.89	0.03	25.0	1.23	0.05
new	32.4	0.80	0.02	36.2	0.75	0.02	38.0	1.13	0.03	36.6	1.12	0.03
par	25.2	1.44	0.06	29.0	0.79	0.03	31.9	1.09	0.03	29.8	0.79	0.03
sil	30.9	0.56	0.02	33.2	0.71	0.02	36.3	1.05	0.03	34.8	0.98	0.03

Table 6.1 – Mean, standard deviation and CoV (PSNR;CBR=256kbps)

Table 6.1 presents results for the PSNR metric when video test sequences are encoded individually, at 256 kbps, using four GOP patterns (IBBP GOP1, IBBP GOP2, IPPP GOP1, and IPPP GOP2). PSNR variability within each video sequence, for the different video programmes, is in most cases quite small. The average value of CoVs, according to GOP pattern, varies between 0.04 (IPPP GOP2 and IBBP GOP2) and 0.05 (IPPP GOP1 and IBBP GOP1). With the exception of football video sequence, all CoV values are bellow 0.10.

It is important to assess the impact of the spatio-temporal video characteristics on the final picture quality. The impact can be measured by analyzing how the picture quality of video programmes varies when the only parameter that changes is the source video, and all the encoding parameters are kept fixed. One way is to measure the relationship between the maximum variation in picture quality for the test video sequences and the minimum value of picture quality for the same set of video sequences. The smallest value of the ratio is obtained for IBBP GOP1 simulations with the value of 59%, and the highest ratio occurs for IPPP GOP1 simulations, with the value of 84%. One can conclude that there is a high range in picture quality variation due to the characteristics of the video sequences. Higher values of picture quality variations are observed in video sequences with IPPP GOP patterns.

The selection of the GOP pattern affects the picture quality of twelve video sequences in a different way. The average value of the picture quality of the video set, for a value of the reference bit rate of 256 kbps, ranges from 29.46 dB (IPPP GOP1) up to 34.21 dB (IBBP GOP1). Thus, the changing GOP pattern may lead to a variation of up to 16% in the value of the

mean picture quality. The impact that the selection of GOP pattern has on the picture quality of a video sequence is not identical in all video sequences. Picture quality ranges from 2% up to 30%.

	IPPP_GOP1			IPPP_GOP2			IBBP_GOP1			IBBP_GOP2		
Video	Mean	Std	CoV	Mean	Std	CoV	Mean	Std	CoV	Mean	Std	CoV
aki	41.8	0.99	0.02	44.1	1.55	0.04	44.4	2.54	0.06	44.1	2.47	0.06
cgd	30.4	0.87	0.03	31.3	1.03	0.03	32.6	1.25	0.04	32.3	1.23	0.04
dea	33.1	0.72	0.02	37.1	1.24	0.03	40.7	2.14	0.05	39.0	2.01	0.05
flg	25.7	0.93	0.04	28.4	1.22	0.04	30.4	1.24	0.04	29.5	1.20	0.04
for	34.4	2.39	0.07	36.2	1.84	0.05	36.6	1.76	0.05	36.2	1.83	0.05
fot	31.4	3.50	0.11	31.9	3.39	0.11	32.1	3.37	0.10	32.0	3.41	0.11
hal	36.9	0.58	0.02	38.5	0.74	0.02	39.7	1.29	0.03	39.2	1.31	0.03
mad	38.3	0.91	0.02	40.2	1.51	0.04	41.1	2.34	0.06	40.8	2.04	0.05
mcl	23.9	1.00	0.04	26.9	0.70	0.03	29.4	1.11	0.04	28.3	0.96	0.03
new	37.0	0.84	0.02	40.1	1.13	0.03	41.1	1.97	0.05	40.2	1.74	0.04
par	29.5	0.50	0.02	33.1	0.95	0.03	35.4	1.55	0.04	33.8	1.06	0.03
sil	33.8	0.54	0.02	36.5	0.97	0.03	40.0	1.88	0.05	38.3	1.70	0.04

Table 6.2 – Mean, standard deviation and CoV (PSNR ;CBR=512kbps)

Table 6.2 presents independent coding simulation results, at 512 kbps, for all the video test sequences. Four GOP patterns were evaluated (IBBP GOP1, IBBP GOP2, IPPP GOP1, and IPPP GOP2). The variability of the PSNR metric at 512 kbps is small, similar to results of simulations made at 256 kbps. The value of the average CoV, aggregated by GOP pattern, varies between 0.04 (IPPP GOP1 and IPPP GOP2) and 0.05 (IBBP GOP1 and IBBP GOP2). Comparing the results obtained at 512 kbps with the results at 256kbps, there is a decrease of the average value of COV for video sequences encoded with the IPPP pattern, and an increase of the mean value of the COVs for sequences encoded with the IBBP pattern. However, looking into the absolute values, CoVs values are low and their variation is small. Again, except for football programmes, CoV values are below 0.10.

To assess the impact of the spatio-temporal characteristics the ratio between the maximum variation of the picture quality among video sequences and the minimum value of picture quality for a video sequence was computed. The smallest value of this ratio occurs for IBBP GOP1 simulations with the value of 51% (59% for 256kbps), and the highest ratio value occurs for IPPP GOP1 simulations, with the value of 75% (84% for 256kbps). Although the measured values of this ratio, for the bit rate 512 kbps, are lower than the values measured for the bit rate 256 kbps, the variation in picture quality remains at high values. Regarding the influence of GOP pattern on picture quality, coded at 512 kbps, the average value of the picture quality varies between 33.02 dB (IPPP GOP1) and 36.97 dB (IBBP GOP1). This corresponds to a

relative variation of 12% of the mean value of the picture quality (16% for 256kbps, 17% for 512 kbps). Analyzing how the picture quality of each video sequence varies with different GOP patterns, a variation from 2% to 23% can be noted.

The variation in picture quality depends strongly on the spatial-temporal characteristics of video sequences. Doubling the bit rate (256 kbps up to 512 kbps) corresponds to an increase, on average, of 10% in the picture quality of video sequences. The variation of picture quality within the encoded video sequences for a certain GOP pattern shows a slight variation compared to the 10% value (8% for IBBP GOP1 up to 12% for IPPP GOP1). A similar impact is seen in the picture quality of video sequences, when the bit rate is doubled without changing the GOP pattern. Hal and Par video sequences present, respectively, the lowest and the highest growth of the average picture quality (6% and 14%). Variation of the picture quality is relatively small within each video programme.

Video sequence's with a longer GOP, display an increase in picture quality, particularly for GOPs that contain a higher number of B pictures. Several video programmes achieve better picture quality results when encoded with the pattern, IBBP GOP1, at 256 kbps, than when encoded at a superior bit rate (512kbps), but the GOP pattern is without B pictures (IPPP GOP1). Thus, the presence of B pictures in a GOP can have a higher effect than other parameters such as bit rate. This is why it is important to evaluate results for both short and long GOPs.

	IPPP_GOP1			IPPP_GOP2			IBBP_GOP1			IBBP_GOP2		
Video	Mean	Std	CoV	Mean	Std	CoV	Mean	Std	CoV	Mean	Std	CoV
aki	48.0	1.34	0.03	56.2	3.12	0.06	61.6	5.97	0.10	58.2	4.69	0.08
cgd	32.4	1.14	0.04	34.0	1.62	0.05	36.0	1.60	0.04	35.4	1.65	0.05
dea	33.6	1.30	0.04	40.4	2.00	0.05	48.7	4.05	0.08	43.9	2.74	0.06
flg	24.7	2.11	0.09	28.5	1.57	0.06	31.8	1.77	0.06	29.9	1.65	0.06
for	36.2	3.08	0.08	39.4	2.77	0.07	40.9	2.74	0.07	39.8	2.66	0.07
fot	33.1	6.08	0.18	34.1	5.99	0.18	34.1	5.76	0.17	33.9	5.56	0.16
hal	40.2	0.93	0.02	46.0	1.01	0.02	50.7	1.97	0.04	48.8	1.91	0.04
mad	42.9	1.39	0.03	48.3	3.32	0.07	53.1	5.48	0.10	50.7	3.85	0.08
mcl	22.9	2.06	0.09	26.6	1.46	0.05	31.3	1.36	0.04	28.6	1.77	0.06
new	39.4	1.43	0.04	46.6	1.99	0.04	50.4	3.02	0.06	47.2	2.71	0.06
par	28.6	1.99	0.07	34.1	1.38	0.04	38.8	2.09	0.05	35.2	1.42	0.04
sil	35.9	0.86	0.02	39.7	1.21	0.03	45.6	2.15	0.05	42.6	1.92	0.04

Table 6.3 – Mean, standard deviation and CoV (PSPNR ;CBR=256kbps)

	IPPP_GOP1			IPPP_GOP2			IBBP_GOP1			IBBP_GOP2		
Video	Mean	Std	CoV	Mean	Std	CoV	Mean	Std	CoV	Mean	Std	CoV
aki	58.2	2.81	0.05	66.5	5.38	0.08	70.2	9.99	0.14	67.8	9.16	0.14
cgd	36.4	1.60	0.04	37.7	1.98	0.05	40.0	2.35	0.06	39.4	2.37	0.06
dea	40.8	1.39	0.03	49.1	2.99	0.06	57.8	6.54	0.11	52.8	5.46	0.10
flg	29.4	1.30	0.04	33.5	1.94	0.06	36.9	2.17	0.06	35.3	2.02	0.06
for	41.8	3.74	0.09	45.3	3.13	0.07	46.3	3.53	0.08	45.3	3.43	0.08
fot	37.9	6.62	0.17	38.9	6.65	0.17	39.1	6.38	0.16	38.9	6.51	0.17
hal	47.7	1.28	0.03	51.3	1.63	0.03	54.2	2.89	0.05	53.1	2.90	0.05
mad	50.8	2.17	0.04	55.7	3.99	0.07	59.6	6.62	0.11	58.5	5.51	0.09
mcl	27.1	1.38	0.05	31.5	1.04	0.03	35.6	2.00	0.06	33.6	1.60	0.05
new	48.8	1.98	0.04	56.6	3.28	0.06	59.8	5.89	0.10	56.8	5.02	0.09
par	35.0	0.85	0.02	41.3	2.08	0.05	46.0	3.40	0.07	42.5	2.29	0.05
sil	40.7	0.95	0.02	46.1	2.03	0.04	55.2	4.88	0.09	50.5	4.04	0.08

Table 6.4 – Mean, standard deviation and CoV (PSPNR ;CBR=512kbps)

Table 6.3 and Table 6.4 present the PSPNR independent coding simulation results, at 256 kbps and 512 kbps respectively, for the complete test video sequences. Four GOP patterns were evaluated (IBBP GOP1, IBBP GOP2, IPPP GOP1, and IPPP GOP2).

The variability of the PSPNR metric presents similar behaviour to the PSNR metric. The mean value CoV in simulations with the bit rate of 256kbps and 512kbps, is 0.065 and 0.074, respectively. The highest values of CoV are observed in video sequences encoded with the pattern IPPP GOP1, and the video sequence with the highest values is the football video sequence, with values of 0.184 (256kbps) and 0.175 (512 kbps). Although the mean value of CoV decreases when the value of the bit rate increases, the variation of the average CoV differs according to the type of GOP pattern. The average value of CoV decreases for video sequences encoded with the IPPP pattern and increases for video sequences encoded with an IBBP GOP pattern. These variations are small in absolute terms. The variation of the mean value of CoV, using PSPNR as the metric to assess picture quality, aggregating the results according to the four patterns of GOP, is higher than that observed in PSNR. In PSNR, the relative variation of the CoV is 17% (256kbps) and 42% (512kbps), while for PSPNR, the variation is 21% (256kbps) and 70% (512kbps). This ratio is computed between the range of CoV variation with the minimum value of CoV, bearing in mind that the values of CoV used in this calculation are the mean values of CoV of video sequences encoded for a particular GOP pattern and bit rate values.

The mean value of PSPNR of video sequences is 39.8 dB (256 kbps) and 46.3 dB (512 kbps). In relative terms, the increase of the bit rate corresponds to an average growth of 17% of the PSPNR. This increase is higher than the increase observed in the PSNR metric (10%).

Computing the average value of PSPNR according to the GOP pattern, the variation is 25% for the video sequences encoded at 256 kbps, and 21% for the video sequences encoded at 512 kbps. The lowest value of PSPNR is obtained for the pattern IPPP GOP1 (34.8 dB / 256 dB and 41.2 kbps / 512 kbps) and the highest value for the pattern IBBP GOP1 (43.6 dB / 50.1 dB and 256 kbps / 512 kbps).

Maintaining the encoding parameters used in the simulations fixed, it can be seen that the differences in picture quality of encoded video sequences are meaningful, particularly in sequences encoded with the IPPP pattern (no images of type B). For example, for the pattern IPPP GOP1, at 256 kbps, the picture quality of the video programmes varies between the value of 22.9 dB (MCL) and the value of 48 dB (Akiyo). Analyzing the results according to the GOP pattern, the video programmes encoded with the IBBP GOP pattern, present the lowest values for the interval at which picture quality varies, and the lowest values for the ratio between this interval and the minimum value quality picture. However, these values are quite high and actually higher than those observed when the PSNR metric is used.

When the bit rate is doubled (IPPP GOP1, 512 kbps), the interval of picture quality variation increases its value (ranges from 27.1 dB, MCL, to 58.2 dB, Akiyo). The ratio between the interval of variation of picture quality and the minimum value of picture quality, while maintaining the coding parameters fixed, is greater than that observed for the PSNR metric (on average, 110% when coded at 256kbps and 115% when coded at 512kbps for PSPNR).

	IPPP_GOP1			IPPP_GOP2			IBBP_GOP1			IBBP_GOP2		
Video	Mean	Std	CoV	Mean	Std	CoV	Mean	Std	CoV	Mean	Std	CoV
aki	72.5	2.21	0.03	82.7	3.11	0.04	85.0	4.96	0.06	83.9	4.39	0.05
cgd	14.1	3.58	0.25	20.6	5.55	0.27	28.7	5.78	0.20	26.8	6.03	0.22
dea	36.1	5.34	0.15	60.2	4.68	0.08	76.8	5.55	0.07	69.1	4.88	0.07
flg	11.0	8.81	0.80	26.4	8.96	0.34	41.0	10.52	0.26	32.4	9.73	0.30
for	34.1	12.76	0.37	45.2	9.54	0.21	50.4	7.97	0.16	47.5	7.78	0.16
fot	14.0	11.85	0.84	16.4	12.54	0.77	17.5	12.18	0.69	16.7	12.04	0.72
hal	52.6	1.69	0.03	61.9	1.34	0.02	67.9	2.35	0.03	66.2	2.12	0.03
mad	54.7	3.80	0.07	68.4	5.97	0.09	74.4	7.68	0.10	71.7	6.07	0.08
mcl	9.3	7.08	0.76	15.7	6.10	0.39	38.4	4.44	0.12	26.0	6.84	0.26
new	48.5	4.61	0.10	68.4	3.04	0.04	75.2	4.47	0.06	69.8	4.71	0.07
par	20.0	6.49	0.32	39.7	3.83	0.10	56.4	5.32	0.09	44.3	3.95	0.09
sil	26.1	3.51	0.13	42.1	4.23	0.10	61.1	6.32	0.10	52.7	5.94	0.11

Table 6.5 – Mean, standard deviation and CoV (SSIM;CBR=256kbps)

	IPPP_GOP1			IPPP_GOP2			IBBP_GOP1			IBBP_GOP2		
Video	Mean	Std	CoV	Mean	Std	CoV	Mean	Std	CoV	Mean	Std	CoV
aki	83.7	2.86	0.03	87.8	3.85	0.04	87.9	5.69	0.06	87.8	5.83	0.07
cgd	30.4	5.12	0.17	36.4	6.81	0.19	46.3	7.18	0.16	44.5	7.46	0.17
dea	60.4	3.45	0.06	76.8	4.44	0.06	86.4	6.13	0.07	82.4	6.31	0.08
flg	29.2	7.98	0.27	48.0	8.76	0.18	58.3	8.36	0.14	54.0	8.58	0.16
for	51.6	12.63	0.24	62.0	7.63	0.12	65.2	6.93	0.11	62.9	6.98	0.11
fot	25.5	14.83	0.58	29.6	14.23	0.48	31.7	15.25	0.48	31.4	14.72	0.47
hal	64.0	1.61	0.03	68.1	2.19	0.03	71.7	3.55	0.05	70.8	3.68	0.05
mad	71.7	3.56	0.05	78.4	5.01	0.06	80.2	7.21	0.09	79.7	6.77	0.08
mcl	16.9	5.31	0.32	36.8	3.25	0.09	55.1	4.76	0.09	48.3	3.90	0.08
new	71.4	3.29	0.05	82.0	3.96	0.05	83.7	6.49	0.08	81.9	5.88	0.07
par	42.1	2.69	0.06	61.9	3.72	0.06	72.3	7.08	0.10	65.4	4.60	0.07
sil	45.7	3.52	0.08	62.2	5.82	0.09	78.6	9.45	0.12	71.8	9.31	0.13

Table 6.6 – Mean, standard deviation and CoV (SSIM;CBR=512kbps)

Table 6.5 and Table 6.6 contain the mean, standard deviation and CoV values of the SSIM picture quality metric, for the independent video coding scenario. Results are presented according to bit rate (256kbps and 512kbps) and GOP patterns (IBBP GOP1, IBBP GOP2, and IPPP GOP1, IPPP GOP2). CoV behaviour differs very much according to the nature of the sequence. Video sequences with low spatial detail and movement present CoV values inferior to 0.10. Video sequences with medium to high levels of spatial and temporal complexity, present very high values of CoV. Examples of these sequences are football, flower garden or mobile and calendar. The average value of the mean CoV is 0.218 for 256 kbps and 0.135 for 512 kbps. CoV variation is affected by GOP pattern. The highest values of CoV variation are obtained for IPPP GOP1 (0.32 and 0.16 for 256 kbps and 512 kbps) and the lowest values for IBBP GOP 1 (16.3% and 12.8% for 256 kbps and 512 kbps). Comparing the three metrics, SSIM metric is the most sensitive to GOP pattern and video characteristics while PSNR the least sensitive. SSIM proposes to be more correlated with the way the human visual system works, thus it is better able to translate how a viewer would perceive picture quality variations.

The average value of SSIM for video sequences encoded at 256kbps is 46.3 and 60.9 for video sequences encoded at 512kbps. By aggregating results by GOP pattern, the average value of SSIM varies between 32.7 and 56.1 for the bit rate of 256kbps, and between 49.4 and 68.1 for the bit rate of 512 kbps. Lowest values are obtained for IPPP GOP1 and higher values for IBBP GOP1. Using these figures, the next step is to compute the ratio between the interval of picture quality variation and the minimum value of picture quality. For encoding at 256kbps, the value of the ratio is 0.71 and for encoding at 512kbps, the value of the ratio is 0.38. The ratio decreases when the bit rate increases. This performance is also observed for PSNR and PSPNR.

Values are high and far exceed those obtained for PSNR and PSPNR. It is thus clear that the SSIM metrics are more sensitive to the choice of the GOP pattern.

Analyzing how the picture quality varies when the coding parameters are fixed (same bit rate and GOP pattern); there are significant differences in the values of SSIM of video sequences. The largest variations occur for IPPP GOP1 [681% and 397% of 256kbps and 512kbps]. The limits of the SSIM values for the IPPP GOP1 pattern are between 9.3 and 72.5 for 256kbps, and between 16.9 and 83.7 for 512kbps.

The selection of GOP pattern has an impact on picture quality that depends also on the spatiotemporal characteristics of each video sequence. For example, Mobile and Calendar video sequence presents the greatest variation (313% for 256kbps and 227% for 512kbps), and the Akiyo video sequence the lowest variation (17% for 256kbps and 5% for 512kbps). The impact of the selection of GOP pattern decreases for higher values of the bit rate. This behaviour is also observed for the PSNR and PSPNR.

Larger variations of picture quality occur in two cases: sequences with high levels of spatial and temporal complexity, and sequences with B pictures in the GOP pattern. The use of B frames originates an important increase of the perceived picture quality of the video sequence. When the bit rate is doubled, the average SSIM obtained in all simulations increases 48%. This rise is greater than the one measured by PSNR or PSPNR (10% for PSNR, and 17% for PSPNR). Moreover, if one compares how growth occurs according to the GOP pattern, it can be found that the video sequences encoded with the IPPP GOP1 pattern presents an increase much higher than with other GOP patterns. In addition, IPPP GOP1 coded video sequences have the lowest values of picture quality regardless of the picture quality metric used.

The goal of this section is to explore joint video coding methods with enhanced rate distortion models based on picture quality metrics. After having analysed independent coding results for the entire set of video sequences, a sub-set was selected to be used in the different joint video coding simulations.

In the first scenario of joint video coding, the number of video programmes that are simultaneously coded is two, and in the second scenario, the number of video programmes is three. For these scenarios, the Akiyo, Foreman and Football video sequences were selected (Figure 5.3). These video sequences represent different levels of spatial and temporal complexity.

The Akiyo video sequence presents the highest levels of picture quality regardless of picture quality metrics (PSNR, SSIM and PSPNR). Akiyo is one of the video sequences where the variation of picture quality is lower due to the choice of the GOP pattern. When varying the

pattern of the GOP, while maintaining a constant bit rate at 256kbps, PSNR can vary up to 4.67dB, PSPNR can vary up to 13.61dB and SSIM can vary up to 12.59. Analyzing the results when the GOP pattern varies, maintaining a constant bit rate at 512kbps, PSNR values can vary up to 2.61dB, PSPNR values may vary up to 12.03 dB and SSIM values can range up to 4.11. In all cases, the variation is smaller for 512kbps than 256kbps. By doubling the bit rate while keeping the GOP pattern, picture quality can improve by up to 10.3% for PSNR, 21.4% for PSPNR and 15.6% for SSIM. The largest increase in picture quality for the Akiyo video sequence is observed when it is coded with the IPPP GOP1 pattern. Regarding the variability of picture quality in the Akiyo video sequence, the observed values are low. For example, the maximum value for the Akiyo sequence CoV is 0.06 (PSNR), 0.07 (SSIM) and 0.14 (PSPNR) for 512kbps and the IBBP GOP pattern.

Analyzing the impact that the GOP pattern has on the picture quality of the simulations for the Foreman video sequence, several observations can be made. One can observe that PSNR can vary up to 2.78 dB (256 kbps), and 2.23 dB (512 kbps), PSPNR can vary up to 4.62 dB (256 kbps), and 4.43 dB (512 kbps), and SSIM can vary up to 16.28 (256 kbps) and 13.54 (512 kbps). Comparing the picture quality results of the Foreman video sequence with the results of the Akiyo video sequence, the degree of variation due to the GOP pattern is higher for SSIM. Average picture quality at reference bit rate of 256kbps can vary up to 17.4% and 47.7% for Akiyo and Foreman, respectively, and at the reference bit rate 512kbps, average picture quality can vary up to 4.9% and 26.2% for Akiyo and Foreman, respectively. As for the results measured by PSNR and PSPNR metrics, the variation in picture quality due to GOP pattern is lower for the Foreman video sequence in comparison with the results obtained for the Akiyo video sequence. At 256kbps, the picture quality for the Akiyo video sequence can vary up to 12% (PSNR) and 28% (PSPNR), and picture quality for the Foreman video sequence can vary up to 9% (PSNR) and 13% (PSPNR). When bit rate is altered to 512 kbps, the picture quality of the Akiyo video sequence can vary up to 6% (PSNR) and 21% (PSPNR), and image quality for the Foreman video sequence can vary up to 6% (PSNR) and 11% (PSPNR). Picture quality increases, for IPPP GOP1, up to 10% for PSNR, 15 % for PSPNR, and 51 % for SSIM, when the bit rate is doubled. The mean increase in the value of picture quality for all GOP patterns is 9% for PSNR, 14% for PSPNR, and 38% for SSIM. The variation of the SSIM metric is higher compared to that observed for the Akiyo sequence (16% to 51% for Akiyo and Foreman). Comparing these results with those obtained for the Akiyo video sequence, the average increase in picture quality, as measured by SSIM and PSNR, is superior at the Foreman video sequence (38% versus 7%), and (9% versus 7%). In the case of the PSPNR results, the average variation in picture quality when the bit rate is doubled is higher for the Akiyo video sequence compared

with the Foreman video sequence (18% versus 14%). The highest values of variation in picture quality when the value of the bit rate is changed are observed for the SSIM.

The coefficient of variation for the Foreman video sequence presents high values for the SSIM metric (the average value is 0.23 for simulations at 256kbps, and the average value is 0.15 for simulations at 512 kbps) and lower values for PSNR and PSPNR (the average CoV is 0.06 for PSNR, and the average value is 0.07 for PSPNR). Typically, the values of CoV for PSNR and SSIM, comparing Foreman and Akiyo video sequences, are higher for simulations with identical encoding parameters. For values of CoV related with PSPNR, this relationship is reversed: the CoV values are usually lower for Foreman compared with Akiyo (the difference is too small at 256kbps, but increases in 512kbps).

The last selected video sequence is the football video sequence. It presents the smallest interval in picture quality within the video sequences when changing the GOP pattern. However, when the bit rate is doubled and the GOP pattern is maintained, there is a high increase in picture quality, particularly in SSIM (the average increase in picture quality is 9.8%, 14.6% and 83.0% respectively for PSNR, PSPNR, and SSIM). The threshold value of the coefficients of variation, for PSPNR and PSNR, is 0.18. This value for SSIM is much higher: 0.47 (512 kbps and IBBP GOP2) and 0.84 (256kbps and GOP1 IPPP). Analysing SSIM results, the CoV values suggest that the variability in picture quality within the video sequence is very high. This behaviour contrasts with the small variability in the picture quality of the Akiyo and Foreman video sequences. The Football video sequence is a sequence with a high degree of temporal and spatial complexity.

In summary, the Akiyo video sequence has the highest picture quality of the set of video sequences. The Football video sequence has one of the lowest picture quality levels and a high level of spatio-temporal complexity, and the Foreman video sequence presents values of picture quality and spatio-temporal complexity in between Akiyo and Football. For the first and second joint video coding simulation scenarios, different combinations of Akiyo, Foreman and Football were assessed according to Table 6.7 and Table 6.8. The Akiyo, Foreman and Football video sequences are represented by letter A, B and C respectively.

Group Name	Sequence Name	Sequence Name
AB	Akiyo	Foreman
AC	Akiyo	Football
BC	Foreman	Football

Table 6.7 – Composition of Group of Two Video Programmes

Group Name	Sequence Name	Sequence Name	Sequence Name
AAB	Akiyo	Akiyo	Foreman
ABB	Akiyo	Foreman	Foreman
ABC	Akiyo	Foreman	Football
BBC	Foreman	Foreman	Football
BCC	Foreman	Football	Football
ACC	Akiyo	Football	Football
AAC	Akiyo	Akiyo	Football

Table 6.8 – Composition of Group of Three Video Programmes

Regarding the third scenario, containing six video programmes, the main criterion was to represent different digital video services. Thus, Akiyo and Football were selected as typical broadcast programmes (news and sports programmes), and Hall as an example of a video surveillance application. To represent video telephone service two video sequences: Mother and Daughter and Silence were selected . The Silence sequence represents a video telephone call using gesture language. Mobile and calendar was also selected to represent a natural movie. Video programmes were selected from the video test sequences defined in the previous Chapter. Each video sequence has a duration of 10 seconds, 30 frames per second, and CIF resolution. Encoding parameters are specified in Table 5.2. For each scenario, it was considered two references bit rate (256 kbps and 512 kbps). Thus in the first joint coding scenario the bandwidth of the transmission channel is 512kbps and 1024kbps, in the second joint coding scenario the bandwidth of the transmission channel is 768kbps and 1536kbps, and in the third joint coding scenario, the bandwidth of the transmission channel is 1536kbps and 3072kbps.

For first two scenarios, simulations were performed using four different GOP Patterns (Table 5.1), and video programmes were grouped according to Table 6.7 and Table 6.8. Simulations in the third scenario were performed for 2 GOP patterns: IBBP GOP1 and IPPP GOP1. This differs from observed results in the literature regarding H.264/AVC as usually results are presented for GOP structures without B-slices. B-slices may use both past and future frames for referencing. As it is needed to buffer future frames, the input delay buffer will increase.

6.3.2 *Joint Video Encoding of Two and Three Programmes*

Picture quality results are presented as the average of the picture quality gain, for the various video programme's combinations, and using the value of the picture quality for the independent coding scenario as a reference. Simulation results for the first two scenarios are displayed in Table 6.9 up to Table 6.16, for the different joint coding methods (Mux Bit, Mux PSNR, Mux PSPNR and Mux SSIM). Picture quality gain is measured for the three picture quality metrics (PSNR, PSPNR and SSIM), at two reference bit rates (256kbps and 512kbps) and with four

different GOP patterns (IBBP GOP1, IBBP GOP2, IPPP GOP1, IPPP GOP2). Complete results for the simulations of the first and second scenarios are available in Annex C.

Bit Rate	Metric	Mux Bit	Mux PSNR	Mux SSIM	Mux PSPNR
256	PSNR	1.55	1.30	1.27	0.44
256	PSPNR	5.35	4.71	5.22	1.56
256	SSIM	11.10	10.48	10.87	3.57
512	PSNR	1.41	1.32	1.36	0.62
512	PSPNR	3.27	2.43	2.87	0.15
512	SSIM	6.30	6.03	5.85	1.45

Table 6.9 – Joint Coding Average Picture Quality Gain (IBBP GOP1; 2SRC)

Bit Rate	Metric	Mux Bit	Mux PSNR	Mux SSIM	Mux PSPNR
256	PSNR	1.91	1.93	1.84	1.18
256	PSPNR	9.44	9.52	9.68	2.10
256	SSIM	13.05	13.46	13.56	3.86
512	PSNR	1.83	1.76	1.75	0.80
512	PSPNR	3.86	3.01	2.94	1.25
512	SSIM	10.72	11.21	11.98	3.40

Table 6.10 – Joint Coding Average Picture Quality Gain (IBBP GOP1; 3SRC)

Table 6.9 and Table 6.10 show performance results of joint coding algorithms (Mux Bit, Mux PSNR, Mux SSIM and Mux PSPNR) for the first and second joint coding scenario when the GOP pattern is IBBP GOP1. All methods present positive average gains. Unless explicitly stated otherwise, average gain will refer to average gain of picture quality. Average gain increases in the second joint coding scenario (jointly coding of three video sources). These are the best results among the tested GOP patterns. IBBP GOP1 is the GOP pattern that contains the higher number of B-slices. B-slices is the most efficient H.264/AVC slice type regarding a bits-quality ratio, and typically require the least amount of bits for a specific picture quality when comparing with I and P slices. The Mux Bit joint coding algorithm records, in general, the best results. However, the Mux PSNR and Mux SSIM joint coding algorithm results are close to the Mux Bit results. The fourth algorithm, Mux PSPNR, presents the lowest picture quality results.

Bit Rate	Metric	Mux Bit	Mux PSNR	Mux SSIM	Mux PSPNR
256	PSNR	0.31	-0.07	-0.04	-0.08
256	PSPNR	-1.92	-2.84	-2.89	-2.88
256	SSIM	1.41	1.35	1.10	1.10
512	PSNR	0.50	0.05	0.16	0.03
512	PSPNR	-1.21	-2.34	-2.11	-2.42
512	SSIM	2.81	1.75	1.88	1.64

Table 6.11 – Joint Coding Average Picture Quality Gain (IBBP GOP2; 2SRC)

Bit Rate	Metric	Mux Bit	Mux PSNR	Mux SSIM	Mux PSPNR
256	PSNR	0.61	0.56	0.57	0.06
256	PSPNR	-1.15	-2.02	-1.98	-1.98
256	SSIM	1.38	1.27	1.35	0.39
512	PSNR	0.43	0.48	0.29	0.03
512	PSPNR	-1.07	-2.54	-1.58	-1.90
512	SSIM	2.01	1.46	1.90	0.53

Table 6.12 – Joint Coding Average Picture Quality Gain (IBBP GOP2; 3SRC)

Table 6.11 and Table 6.12 present IBBP GOP2 results for the first and second joint encoding scenario. Picture quality measured by PSNR presents low picture quality gains. SSIM average gain is repeatedly positive regarding the different combinations. On the other hand, PSPNR results are always negative. Results improve when the bit rate increases. In general, Mux Bit, Mux PSPNR and Mux SSIM present similar results. Mux PSPNR results are below the other three joint video coding strategies. This type of behaviour of the different joint coding algorithms is also observable in IBBP GOP1 results.

Bit Rate	Metric	Mux Bit	Mux PSNR	Mux SSIM	Mux PSPNR
256	PSNR	-0.19	-0.21	-0.22	-0.65
256	PSPNR	-0.41	-0.44	-0.45	-1.34
256	SSIM	0.69	0.62	0.67	-0.03
512	PSNR	0.05	0.02	0.00	-0.25
512	PSPNR	-0.45	-0.54	-0.59	-1.42
512	SSIM	2.53	2.52	2.56	1.33

Table 6.13 – Joint Coding Average Picture Quality Gain (IPPP GOP1; 2SRC)

Bit Rate	Metric	Mux Bit	Mux PSNR	Mux SSIM	Mux PSPNR
256	PSNR	-0.01	-0.06	-0.03	-0.41
256	PSPNR	-0.29	-0.40	-0.36	-1.06
256	SSIM	0.72	0.64	0.70	-0.91
512	PSNR	0.17	0.10	0.16	-0.07
512	PSPNR	-0.08	-0.16	-0.20	-0.87
512	SSIM	1.88	1.86	1.89	1.11

Table 6.14 – Joint Coding Average Picture Quality Gain (IPPP GOP1; 3SRC)

Table 6.13 and Table 6.14 show IPPP GOP1 picture quality values for the average gain. In general, this is the least favourable scenario regarding the mean value of joint coding gain. It is important to observe that looking at the independent coding results, IPPP GOP1, corresponds to the lowest picture quality values when comparing with other GOP patterns and the range of variation is the smallest.

If IBBP GOP1 is the GOP pattern that contains the highest number of B frames, IPPP GOP1 contains the highest number of Intra frames. Intra frames usually present the worst bits/quality

ratio demanding more bits than P and B frames in order to achieve a certain level of quality. Results improve with the increase of the reference bit rate from 256 kbps to 512 kbps. Once again, the lowest results are obtained by Mux PSPNR.

Bit Rate	Metric	Mux Bit	Mux PSNR	Mux SSIM	Mux PSPNR
256	PSNR	0.10	0.00	0.00	-0.01
256	PSPNR	-0.08	-0.32	-0.29	-0.33
256	SSIM	1.67	1.56	1.57	1.56
512	PSNR	0.31	0.28	0.28	0.28
512	PSPNR	0.02	-0.22	-0.21	-0.23
512	SSIM	3.74	3.85	3.80	3.83

Table 6.15 – Joint Coding Average Picture Quality Gain (IPPP GOP2; 2SRC)

Bit Rate	Metric	Mux Bit	Mux PSNR	Mux SSIM	Mux PSPNR
256	PSNR	0.21	0.13	0.15	0.12
256	PSPNR	0.00	-0.21	-0.18	-0.24
256	SSIM	1.41	1.23	1.29	1.27
512	PSNR	0.26	0.24	0.24	0.24
512	PSPNR	0.04	-0.12	-0.12	-0.13
512	SSIM	3.00	3.15	3.14	3.13

Table 6.16 – Joint Coding Average Picture Quality Gain (IPPP GOP2; 3SRC)

Table 6.15 and Table 6.16 present results for the fourth and last GOP pattern: IPPP GOP2. Average gain is near zero for PSNR but from 1.23 up to 3.85 for SSIM. Average gain achieves its lowest results when the PSPNR picture quality metric is used. Mean picture quality results improve when bit rate increases from 256 kbps to 512 kbps.

To provide a better understanding of Mux Bit, Mux PSNR, Mux SSIM and Mux PSPNR, Figure C.1 up to Figure C.8, in Annex C, show picture quality, PSNR and SSIM, for the first 150 frames of the video programme of Akiyo, Foreman and Football, for the first and second joint coding scenarios. GOP pattern is IPPP GOP2 and bit rate reference is 256 kbps. In all the figures, the independent coding scenario (CBR) results are also included as a reference.

In all the different combination cases, for the first and second scenario, the picture quality of Akiyo decreases when combined with the Foreman and Football video sequences and the picture quality of football always increases when combined with Akiyo or Foreman. This is expected, as Akiyo is the sequence with higher quality picture values in independent coding and Football the video sequence with the lower picture quality results in independent coding.

Results from using joint rate control show that the picture quality gap between different video programmes decreases, and a higher uniform picture quality among sequences is achieved. Furthermore, the picture quality variation within a sequence has a tendency to be much smoother in joint coding than in independent coding. The degree to which picture quality can be

uniform depends on the nature of the video sequence. For example, picture quality in the football video sequence, especially when it is measured using SSIM, presents a high variability when coded individually. Psychological studies advise that human observers favour a video sequence with consistent visual quality to a video sequence with varying visual quality ([95]). Thus reducing variability of picture quality within a video sequence can increases perceived picture quality.

6.3.3 *Joint Video Encoding of Six Programmes*

The first two scenarios provide some important clues that are further exploited in the third scenario. From the first and the second joint coding scenarios it is possible to conclude that sequence characteristics, GOP pattern and reference bit rate play an important role in joint coding. Simulation results such as the picture quality gain depend on the picture quality metric. Thus, in some cases, according to one picture quality metric, the value of the average gain is positive and according to another picture quality metric, the value of the average gain is negative. Nevertheless, in both cases the lower picture quality is always increased. Results also show that the use of perceptual metrics, particularly SSIM, provide results equivalent to the results obtained by traditional approaches such as is the case with Mux Bit.

Before performing an analysis of the simulation results of the third joint video coding scenario, independent coding results are presented (Table 6.17 and Table 6.18). To ensure a more complete study of the simulation results, it is necessary to estimate the values of certain parameters in order to complement the description of the data set. The parameters selected should allow the assessment of how similar the different observations are (measures of central tendency) and how different the different observations are (measures of dispersion). Max, Mean, Min and Stdev values correspond respectively to the maximum, the average, the minimum, and the standard deviation values of the picture quality of the six video programmes. Range is the difference between the maximum and the minimum picture quality value of the six video programmes and CoV is the Coefficient of Variation of the six video programmes as defined in Equation (5.50).

	IPPP GOP1			IPPP GOP2			IBBP GOP1			IBBP GOP2		
	PSNR	PSPNR	SSIM	PSNR	PSPNR	SSIM	PSNR	PSPNR	SSIM	PSNR	PSPNR	SSIM
Max	37.90	47.95	72.45	41.21	56.15	82.67	42.57	61.56	85.04	41.70	58.24	83.89
Mean	30.98	37.14	38.19	33.45	41.83	47.86	35.33	46.06	57.40	34.37	43.80	52.86
Min	20.63	22.86	9.28	23.57	26.65	15.74	26.78	31.28	17.53	25.00	28.64	16.68
Stdev	5.99	8.73	25.37	6.30	10.58	27.88	6.09	11.61	25.02	6.22	11.06	26.55
CoV	0.19	0.23	0.64	0.18	0.25	0.54	0.16	0.24	0.39	0.17	0.24	0.45
Range	17.3	25.1	63.2	17.6	29.5	66.9	15.8	30.3	67.5	16.7	29.6	67.2

Table 6.17 – Statistical results of independent video coding (6SRC; 256 kbps)

	IPPP GOP1			IPPP GOP2			IBBP GOP1			IBBP GOP2		
	PSNR	PSPNR	SSIM	PSNR	PSPNR	SSIM	PSNR	PSPNR	SSIM	PSNR	PSPNR	SSIM
Max	41.80	58.21	83.75	44.06	66.51	87.84	44.41	70.24	87.86	44.08	67.84	87.76
Mean	34.08	43.74	51.24	36.41	48.32	60.50	37.91	52.31	67.55	37.09	50.41	64.96
Min	23.91	27.11	16.85	26.89	31.47	29.64	29.43	35.57	31.75	28.27	33.59	31.39
stdev	5.68	10.90	26.52	5.51	12.44	22.98	5.09	12.96	20.73	5.25	12.58	21.07
CoV	0.16	0.25	0.48	0.15	0.26	0.35	0.13	0.24	0.28	0.14	0.24	0.30
Range	17.9	31.1	66.9	17.2	35.0	58.2	15.0	34.7	56.1	15.8	34.3	56.4

Table 6.18 – Statistical results of independent video coding (6SRC; 512 kbps)

Analysing Table 6.17 and Table 6.18, the highest mean picture quality results are obtained for IBBP GOP1 and the lowest picture quality are obtained for IPPP GOP1. It is to be noted that IBBP GOP1 and IPPP GOP1 present respectively the lowest and the highest picture quality variation when the reference bit rate changes from 256kbps to 512kbps. These two GOP patterns were selected for the third joint coding scenario.

For IBBP GOP1 results, when the value of the bit rate is doubled, the mean value of the picture quality of the video sequences increases its value (7% for PSNR, 14% for PSPNR and 18% for SSIM). The value of the range, in IBBP GOP1, is reduced 5% for PSNR and 17% for SSIM, and the value of the range is increased 15% for PSPNR. The standard deviation displays a similar behaviour regarding the performance of the range parameter (when bit rate increases, stdev decreases (6% for PSNR and 17% for SSIM, and stdev increases 12% for PSPNR). Regarding CoV, its value decreases for all the three picture quality metrics when bit rate increases (12% for PSNR, 2% for PSPNR and 29% for SSIM).

An increment in the bit rate has a distinct impact on the picture quality of the different video programmes. In particular, in the parameters Min and Max. For example, when doubling the bit rate, the minimum value increases 10% for PSNR, 14% for PSPNR, and 81% for SSIM, and the maximum value increases 4% for PSNR, 14% for PSPNR, and 3% for SSIM. When bit rate is doubled, the amplitude of parameters variation, such as Max, Min or CoV, is higher for IPPP GOP1 simulations results when compared with IBBP GOP1 simulation results. The variability

within sequences and between sequences decreases when the value of bit rate is doubled. The only statisrical parameter that decreases its value when the bit rate increases its value is CoV. All the remaining parameters increase their values when bit rate is increased.

The greatest increase of the mean value of the picture quality among the six video sequences, when the bit rate is doubled, occurs in simulations with the pattern of IPPP GOP1 (11% for PSNR, 18% and 34% for PSPNR SSIM). However, when analysing the absolute values, the mean values of picture quality are the lowest values that were obtained (the highest value occurs with IBBP GOP1). As for the maximum values, considering both reference bit rates of 256kbps and 512kbps, IPPP GOP1 maximum values are the lowest values compared with the values obtained in simulations with other patterns of GOP. However, when compared with the relative growth of its value when bit rate is doubled, they present the highest relative growth (10% for PSNR, 21% for PSPNR, and 16% for SSIM). As for the minimum value, IBBP GOP1 presents the lowest values and one of the highest increases. However, the difference is not as substantial as in the case of the maximum (16% for PSNR, 19% for PSPNR, and 82% for SSIM). The lowest values for the parameters max, mean and min were obtained for simulations with the pattern IPPP GOP1, at both 256 kbps and 512 kbps. The highest variation of the parameters max, mean and min occurs for SSIM and the lowest variation for PSNR. As for range and stdev, the highest values of variations take place for PSPNR and the lowest values of the variations occur for PSNR.

We will now look at the individual results from each video sequence. One of the video sequences that presents a high level of spatio-temporal complexity and whose results show that it is a difficult sequence to encode is the Mobile and Calendar sequence. For example, the value of SSIM of the Mobile and Calendar video sequence encoded at 1024kbps with IPPP GOP1, is roughly half the value of the SSIM value for the Akiyo video sequence when encoded at 256kbps and IPPP GOP1.

	PSNR		PSPNR		SSIM	
Video	IBBP GOP1	IPPP GOP1	IBBP GOP1	IPPP GOP1	IBBP GOP1	IPPP GOP1
aki	1.84	3.90	8.68	10.26	2.82	11.30
fot	2.91	2.87	5.08	4.85	14.22	11.45
hal	1.44	3.67	3.44	7.55	3.81	11.39
mad	2.17	3.60	6.43	7.87	5.82	17.05
mcl	2.65	3.27	4.29	4.25	16.77	12.57
sil	3.76	2.89	9.63	4.80	17.47	19.55

Table 6.19 – Picture Quality Increment when bit rate is doubled for the independent coding

Table 6.19 shows the increase in picture quality for each video sequence, when the bit rate is doubled, in the independent coding scenario. From Table 6.19 it can be seen that the extent of

the increase in image quality depends on the sequence characteristics and the GOP pattern. In sequences with low spatial and temporal resolution, such as Akiyo, Hall or Mother and Daughter, the increment in picture quality, in general, is higher for IPPP GOP1 compared with IBBP GOP1. In sequences with medium to high spatial and temporal resolution, such as Football or Mobile and Calendar, the increment in picture quality, in general, is higher for IBBP GOP1 compared with IPPP GOP1. From Table 6.19 the difference between the maximum value and the minimum value of the increase in image quality according to GOP pattern can also be analyzed. Hence, it can be seen that the variation interval is smaller for IPPP GOP1 (PSNR: 1.0 dB, PSPNR: 6.0 dB, SSIM: 8.3) when compared to IBBP GOP1 (PSNR: 2.3 dB, PSPNR: 6.2 dB, SSIM: 14.7). The difference of the interval of variation is particularly large for the SSIM picture quality metric.

After this brief analysis of the simulation results when the six video programmes are encoded independently, an analysis of results for the third joint video coding scenario follows. As mentioned earlier, in the third scenario corresponding to the joint encoding of six video sources, simulations were performed for two different reference bit rates (256 kbps and 512 kbps) and two GOP patterns (IBBP GOP1 and IPPP GOP1). Thus, each video sequence was encoded four times with different coding settings for each joint video coding algorithm. From Table 6.20 to Table 6.23, several statistical parameters are displayed (Max, Average, Min, Stdev, CoV and Range) concerning the simulation results of the third coding scenario. Individual results of the simulations are available in Annex C (Table C.9 to Table C.12).

	Mux Bit			Mux PSNR			Mux SSIM			Mux PSPNR		
	PSNR	PSPNR	SSIM	PSNR	PSPNR	SSIM	PSNR	PSPNR	SSIM	PSNR	PSPNR	SSIM
Max	40.59	57.98	77.65	41.28	60.06	80.50	40.18	59.03	75.35	40.77	59.72	77.21
Mean	35.33	45.92	58.29	35.47	46.18	58.13	35.35	46.35	58.55	35.43	46.38	58.47
Min	29.09	34.00	17.03	27.75	32.79	21.12	28.21	33.23	29.24	28.20	33.57	23.93
stdev	4.50	8.81	21.25	5.18	10.51	21.85	4.74	9.52	17.66	4.97	10.01	19.92
CoV	0.13	0.19	0.36	0.15	0.23	0.38	0.13	0.21	0.30	0.14	0.22	0.34
Range	11.50	23.98	60.62	13.53	27.27	59.38	11.97	25.80	46.11	12.57	26.15	53.28

Table 6.20 – Max, Mean, Min, stdev, CoV, Range (6SRC; IBBP GOP1; 256 kbps)

	Mux Bit			Mux PSNR			Mux SSIM			Mux PSPNR		
	PSNR	PSPNR	SSIM	PSNR	PSPNR	SSIM	PSNR	PSPNR	SSIM	PSNR	PSPNR	SSIM
Max	41.46	60.31	84.92	43.61	66.40	86.72	43.61	66.41	86.73	42.97	63.88	85.71
Mean	37.54	50.43	67.88	37.96	52.11	68.30	37.82	51.76	69.24	37.85	51.73	69.07
Min	31.58	38.97	27.56	31.21	37.05	36.42	29.91	36.01	44.09	30.44	37.45	40.53
stdev	3.91	8.11	20.50	4.82	10.99	17.65	5.09	11.05	14.98	4.78	10.31	16.06
CoV	0.10	0.16	0.30	0.13	0.21	0.26	0.13	0.21	0.22	0.13	0.20	0.23
Range	9.88	21.34	57.36	12.40	29.35	50.30	13.70	30.40	42.64	12.53	26.43	45.18

Table 6.21 – Max, Mean, Min, stdev, CoV, Range (6SRC; IBBP GOP1; 512 kbps)

	Mux Bit			Mux PSNR			Mux SSIM			Mux PSPNR		
	PSNR	PSPNR	SSIM	PSNR	PSPNR	SSIM	PSNR	PSPNR	SSIM	PSNR	PSPNR	SSIM
Max	34.25	44.88	66.93	37.12	46.47	70.51	36.33	47.21	69.70	36.74	45.16	70.15
Mean	30.11	36.86	37.47	31.06	37.23	37.68	31.02	37.12	38.11	30.99	37.06	37.65
Min	24.30	28.96	13.99	21.90	24.45	8.37	21.77	24.26	7.63	22.42	25.81	7.20
stdev	3.46	5.45	20.21	5.24	7.69	24.59	5.09	7.76	23.86	4.95	6.82	24.56
CoV	0.11	0.15	0.54	0.17	0.21	0.65	0.16	0.21	0.63	0.16	0.18	0.65
Range	9.95	15.92	52.94	15.22	22.02	62.14	14.56	22.95	62.07	14.32	19.35	62.95

Table 6.22 – Max, Mean, Min, stdev, CoV, Range (6SRC; IPPP GOP1; 256 kbps)

	Mux Bit			Mux PSNR			Mux SSIM			Mux PSPNR		
	PSNR	PSPNR	SSIM	PSNR	PSPNR	SSIM	PSNR	PSPNR	SSIM	PSNR	PSPNR	SSIM
Max	37.76	47.57	71.55	40.81	56.92	80.63	40.52	56.56	79.78	40.05	53.73	78.69
Mean	33.24	41.21	49.78	34.08	43.50	51.21	34.08	43.45	52.61	34.02	42.92	49.75
Min	27.59	34.38	24.98	24.78	28.44	19.77	24.54	28.34	21.99	25.12	29.27	21.97
stdev	3.66	5.01	16.94	5.52	9.89	23.66	5.52	9.72	22.32	5.18	8.65	22.66
CoV	0.11	0.12	0.34	0.16	0.23	0.46	0.16	0.22	0.42	0.15	0.20	0.46
Range	10.17	13.19	46.57	16.03	28.48	60.86	15.98	28.22	57.79	14.93	24.46	56.72

Table 6.23 – Max, Mean, Min, stdev, CoV, Range (6SRC; IPPP GOP1; 512 kbps)

Usually the highest Mean parameter results are obtained by the algorithms Mux PSNR and Mux SSIM. The lowest results are obtained with the algorithm Mux Bit. The Mean parameter presents minor variations for the different algorithms used to jointly encode the video sequences (up to 3.2% for PSNR, 5.5% for PSPNR, and 5.8% for SSIM). The difference between Mean values is higher in simulations using the IPPP GOP1 pattern compared to simulations using IBBP GOP1 pattern. In addition, the relative difference between the results of the Mean parameter, due to the use of different encoding schemes, is smaller when using the IBBP GOP1 pattern than when using the IPPP GOP1 pattern.

Regarding the parameters associated with the boundary values of picture quality, Min and Max; their relative variation is higher than in the case of the Mean parameter. The highest values of the Max parameter are achieved with the algorithms Mux PSNR (256kbps) and Mux SSIM (512kbps). The lowest results of the Max parameter are obtained with the algorithm Mux Bit. The differences between the values for the parameter Max in the different simulations is below 20%. The largest differences occur for PSPNR (19.7%) and the smallest for PSNR (8.4%). Simulations with the IBBP GOP1 pattern present higher Max parameter values compared with simulations with the IPPP GOP1 pattern.

The best results of the Min parameter are obtained mainly with the algorithm Mux Bit and the lowest values of the Min parameter with the algorithms Mux PSNR and Mux SSIM. The relative variation of the value of the Min parameter due to the joint coding algorithms is greater

than the variation of the values of the Mean and the Max parameters. The greatest variation of the value of the parameter Min takes place when the SSIM metric is used, and the lowest variation occurs with PSNR as the picture quality metric. The parameter Min presents higher values for simulations with the pattern IBBP GOP1 than in simulations with IPPP GOP1.

An analysis of the parameters Min, Mean and Max makes it possible to identify similarities in the way the parameters vary. The highest values of the parameters are always obtained for simulations with IBBP GOP1 pattern. On the other hand, the variation of the parameters' value, due to the different joint coding algorithms, is higher for simulations with the IPPP GOP1 pattern. Analyzing the picture quality metrics, regardless of bit rate or the pattern of the GOP, PSNR always shows the lowest variation among its values for the different joint coding algorithms. At the other extreme, SSIM displays the highest variation among its values for the simulations resulting from the different joint coding algorithms. SSIM is generally associated with perceptual image quality. Thus the greater variability of picture quality due to the use of different strategies for the joint coding of video sources, measured by the SSIM metric, can be understood as a signal that the perceived image quality can suffer variations greater than what would be expected if only PSNR was used as picture quality metric. In the next section, the results of a subjective assessment test with a panel of human viewers are presented. The goal of next section is to study how the video simulation results, in the current joint coding scenario, are perceived by an observer. A second goal is to study how the picture quality metrics (PSNR, PSPNR, and SSIM) translate the opinion of an observer regarding the picture quality of the set of video programmes. Video subjective quality tests are too costly in terms of time and human resources, so the tests were limited only to the current joint coding scenario.

Several observations can be made on the parameters that measure how different the different observations are (stdev, CoV and range). In general, the lowest values of the parameters stdev, CoV and Range, are obtained with the Mux Bit joint video coding algorithm. The variation of the values of the stdev parameter, the CoV parameter and the range parameter is higher in video sequences encoded with the IPPP GOP1 pattern for the image quality metrics PSNR and PSPNR. Regarding the SSIM metric, the relationship of GOP pattern is reversed (the relative difference is generally higher for video sequences encoded with the IBBP GOP1 pattern). The value of the CoV parameter differs considerably according to the selected image quality metric.

When PSNR is used to evaluate the quality of the video sequences, the values of the CoV parameter are always below 0.17, for both the two different patterns of GOP and for the two reference bit rates. These low values indicate a low variability of image quality within the video sequences (the mean value of CoV for PSNR is 0.140). The mean value of the CoV parameter of PSPNR, during the simulations, is 0.197. Typically, the values of the CoV for SSIM are

higher than PSNR and PSPNR values: the mean value of the Coefficient of Variation of SSIM results is 0.409 and the highest value goes up to 0.65. The mean values of the CoV parameter for SSIM show a high variability of perceived image quality. These observations about the relationship between the variability of image quality and choosing a metric to assess the image quality are in line with earlier observations (analysis of Max, Mean and Min parameters).

One of the aims of this section is to compare the results obtained using algorithms for joint coding of video sources with the results obtained for independent coding of video sources. Table 6.17 and Table 6.18 contain results of simulations for independent coding, and simulation results for joint coding are available from Table 6.20 to Table 6.23. To allow a better reading, the difference between the value of each parameter obtained for joint coding and the value of the same parameter obtained for independent coding was computed (Table 6.24 to Table 6.27).

	Mux Bit			Mux PSNR			Mux SSIM			Mux PSPNR		
	PSNR	PSPNR	SSIM	PSNR	PSPNR	SSIM	PSNR	PSPNR	SSIM	PSNR	PSPNR	SSIM
vMax	-1.98	-3.58	-7.39	-1.29	-1.50	-4.54	-2.39	-2.53	-9.69	-1.80	-1.84	-7.83
vMean	0.00	-0.14	0.89	0.14	0.12	0.73	0.02	0.29	1.15	0.10	0.32	1.07
vMin	2.31	2.72	-0.50	0.97	1.51	3.59	1.43	1.95	11.71	1.42	2.29	6.40
vstdev	-1.59	-2.80	-3.77	-0.91	-1.10	-3.17	-1.35	-2.09	-7.36	-1.12	-1.60	-5.10
vCoV	-0.03	-0.05	-0.03	-0.01	-0.01	-0.01	-0.03	-0.03	-0.09	-0.02	-0.02	-0.05
vRange	-4.30	-6.32	-6.88	-2.27	-3.03	-8.12	-3.83	-4.50	-21.39	-3.23	-4.15	-14.22

Table 6.24 – Difference between joint coding and independent coding simulation results (6SRC; IBBP GOP1; 256 kbps)

	Mux Bit			Mux PSNR			Mux SSIM			Mux PSPNR		
	PSNR	PSPNR	SSIM	PSNR	PSPNR	SSIM	PSNR	PSPNR	SSIM	PSNR	PSPNR	SSIM
vMax	-2.95	-9.93	-2.94	-0.80	-3.84	-1.14	-0.80	-3.83	-1.13	-1.44	-6.36	-2.15
vMean	-0.37	-1.88	0.33	0.05	-0.20	0.75	-0.09	-0.55	1.69	-0.06	-0.58	1.52
vMin	2.15	3.40	-4.19	1.78	1.48	4.67	0.48	0.44	12.34	1.01	1.88	8.78
vstdev	-1.18	-4.85	-0.23	-0.27	-1.97	-3.08	0.00	-1.91	-5.75	-0.31	-2.65	-4.67
vCoV	-0.03	-0.08	0.02	0.00	-0.03	-0.02	0.00	-0.03	-0.06	0.00	-0.04	-0.05
vRange	-5.12	-13.36	1.26	-2.60	-5.35	-5.80	-1.30	-4.30	-13.46	-2.47	-8.27	-10.92

Table 6.25 – Difference between joint coding and independent coding simulation results (6SRC;IBBP GOP1; 512kbps)

	Mux Bit			Mux PSNR			Mux SSIM			Mux PSPNR		
	PSNR	PSPNR	SSIM	PSNR	PSPNR	SSIM	PSNR	PSPNR	SSIM	PSNR	PSPNR	SSIM
vMax	-3.65	-3.07	-5.52	-0.78	-1.48	-1.94	-1.57	-0.74	-2.75	-1.16	-2.79	-2.30
vMean	-0.87	-0.28	-0.72	0.08	0.09	-0.51	0.04	-0.02	-0.08	0.01	-0.08	-0.54
vMin	3.67	6.10	4.71	1.27	1.59	-0.91	1.14	1.40	-1.65	1.79	2.95	-2.08
vstdev	-2.53	-3.28	-5.16	-0.75	-1.04	-0.78	-0.90	-0.97	-1.51	-1.04	-1.91	-0.81
vCoV	-0.08	-0.08	-0.10	-0.02	-0.02	0.01	-0.03	-0.02	-0.01	-0.03	-0.05	0.01
vRange	-7.35	-9.18	-10.26	-2.08	-3.08	-1.06	-2.74	-2.15	-1.13	-2.98	-5.75	-0.25

Table 6.26 – Difference between joint coding and independent coding simulation results (6SRC;IPPP GOP1; 256 kbps)

	Mux Bit			Mux PSNR			Mux SSIM			Mux PSPNR		
	PSNR	PSPNR	SSIM	PSNR	PSPNR	SSIM	PSNR	PSPNR	SSIM	PSNR	PSPNR	SSIM
vMax	-4.04	-10.64	-12.20	-0.99	-1.29	-3.12	-1.28	-1.65	-3.97	-1.75	-4.48	-5.06
vMean	-0.84	-2.53	-1.46	0.00	-0.25	-0.03	0.00	-0.30	1.37	-0.06	-0.82	-1.49
vMin	3.68	7.27	8.13	0.87	1.33	2.92	0.63	1.23	5.14	1.21	2.16	5.12
vstdev	-2.02	-5.89	-9.58	-0.16	-1.01	-2.86	-0.16	-1.18	-4.20	-0.50	-2.25	-3.86
vCoV	-0.05	-0.13	-0.14	0.00	-0.02	-0.02	0.00	-0.03	-0.06	-0.01	-0.05	-0.02
vRange	-7.73	-17.91	-20.33	-1.87	-2.62	-6.04	-1.92	-2.88	-9.11	-2.97	-6.64	-10.18

Table 6.27 – Difference between joint coding and independent coding simulation results (6SRC;IPPP GOP1; 512kbps)

From Table 6.24 to Table 6.27, all the parameters start with the prefix “v”. When placed in front of a parameter, the prefix "v" indicates the variation of the parameter value between the results for joint coding of video sources and the results obtained for independent coding of video sources. The simulation results show an increase in the value of the parameter that measures the minimum value of the image quality, and the reduction of the value of the range parameter for all the joint coding experiments. Thus, the results show that through a process of joint coding of video sequences, the available bandwidth can be reallocated between the different video programmes resulting in a uniform "distribution" of the image quality for the set of test video sequences. Additionally, there is also an increase in the minimum value of the image quality concerning the set of test video sequences.

The amplitude of the variation of the Mean parameter is reduced, and depending on the algorithm, picture quality metrics, and coding settings can be positive or negative. When the results are analyzed according to the pattern of GOP, the highest values of the Mean parameter variation are obtained for the pattern of IBBP GOP1. The use of the B-slices frame provides a more efficient way to encode a frame compared to other types of images, such as the I- or P-slices. Therefore, when the bandwidth is changed, the slices encoded as B-slices offer a better adaptation to the new value of the bandwidth. This fact is important because most of the existing studies reported in the literature do not analyze the impact of B-slices.

It is clear that the performance levels of each joint coding algorithm depend on the selection of the image quality metric. For example, comparing the different strategies of joint coding, the algorithm Mux SSIM has the best results for the SSIM metric, but poor results for the PSNR metric and the PSPNR metric.

The Mux SSIM algorithm has the best set of results, for the parameters vMin, vMean and vRange, for the IBBP GOP1 pattern, when the SSIM metric is used to evaluate the image quality. Regarding the IPPP GOP1 pattern, the Mux SSIM algorithm has the best results for the vMean parameter, and among the best results for the vMin parameter and the vRange parameter, when the SSIM metric is used to evaluate image quality. For both GOP patterns, the performance of the Mux SSIM algorithm improves when the bit rate is increased, when the SSIM metric is used to evaluate the image quality. If the results of simulations are analysed using the values of PSNR or the values of PSPNR as an image quality indicator, then the analysis differs from the previous one that uses SSIM values. The Mux SSIM performance, when PSNR and PSPNR are used to assess image quality, is lower and decreases when the bit rate is increased. As the algorithm is based on the use of the SSIM metric, this result is expected. The association between SSIM and the perception of image quality has the potential to help ensure a better image quality that is perceived by an observer. Subjective quality assessment tests should allow the validity of this hypothesis to be assessed.

In general, the Mux PSNR algorithm presents the best results for the vMax and vMean parameters when PSNR or PSPNR are used to evaluate image quality. However, for the vMean parameter, the differences are rather small between the different joint coding algorithms. For the Mux PSNR algorithm, the values of the vMean parameter are low, near to zero, for simulations with the IPPP GOP1 pattern. In simulations using the Mux PSNR algorithm but with the IBBP GOP1 pattern, the value of the vMean parameter is higher, and increases significantly when the bit rate is doubled. In addition, the Mux PSNR algorithm presents the lowest set of values of the dispersion measuring parameters (vStdev, vCoV and vRange). The results of the Mux PSNR algorithm when compared with the remaining joint coding algorithms show a slight increase in the average value of the image quality of the video sequences, and a greater variability in image quality between the video sequences.

The Mux PSPNR algorithm presents results that are half way between the best and worst results. The Mux PSPNR algorithm performs better in simulations with IBBP GOP1 than in simulations with the IPPP GOP1 pattern. For the IPPP GOP1 pattern, the results of the vMean parameter are in third position (in a group of four algorithms). However, for IPPP GOP1, the Mux PSPNR algorithm displays good results for the dispersion measuring parameters, when PSNR or PSPNR are used to evaluate image quality. For the IBBP GOP1 pattern, one can

observe that the performance improves in comparison with other algorithms for the different parameters, particularly for the vMean parameter. In general, the Mux PSPNR algorithm achieves better performances at higher bit rates.

The Mux Bit algorithm has the worst results in relation to the vMax and vMean parameters and the best results regarding the vMin parameters and the dispersion measuring parameters (vstdev, vCoV, and vRange). In order to better understand the behavior of the Mux Bit algorithm, a closer analysis of the rate-distortion performance of each video sequence is needed. The Mux Bit algorithm is performed in a two pass encoding. The first pass estimates the complexity of video sequences using a first set of fixed quantisation parameters. Complexity in a video sequence is defined as the product of the effective number of bits used to encode an image and the average quantiser parameter over the whole picture. The second pass encodes the video sequences by redistributing the bandwidth among the video sequences according to the complexity of each video sequence obtained in the first pass. As the quantisation parameter is the same for all video sequences then the complexity depends mainly on the effective number of bits used to encode an image in the first pass. Analyzing the amount of bits generated during the first pass, for different video sequences, there is a wide discrepancy between the Mobile and Calendar video sequence and the remaining video sequences. Mobile and Calendar is the sequence that in the first pass requires more bits to be encoded, followed by the video sequences Football and Silence. The sum of the number of bits needed to encode these three video sequences, in the first pass, goes up to 72% of the total number of bits that it is necessary to encode for all six video sequences. Mobile and Calendar generates in the first pass 2.57 times more bits than the Football video sequence and 2.81 times more bits than the Silence video sequence. Due to the combination of the criteria for allocation of bandwidth in the Mux Bit algorithm and the characteristics of the set of video sequences, one can see that the Mobile and Calendar video sequence is the only sequence whose value of bandwidth is increased compared with the independent coding scenario. The Mobile and Calendar video sequence globally present the lowest values of image quality for all the three image quality metrics (PSNR, PSPNR and SSIM). Thus, as Mobile and Calendar receives more bits than in the remaining joint coding methods, the vMin and vRange parameters are strongly improved compared with the remaining joint coding algorithms.

Looking at Table 6.19 it can be seen how differently the image quality of the video sequence varies with bit rate due to the GOP pattern. For example, the Akiyo sequence varies for the PSNR metric 1.84dB (IBBP GOP1) and 3.9dB (IPPP GOP1) when bit rate is doubled, and, for the same coding settings, Mobile and Calendar varies 2.65dB (IBBP GOP1) and 3.27dB (IPPP GOP1). The Mux Bit algorithm can obtain better results for the vMean parameter when the

video sequences with the highest levels of complexity correspond to sequences with the highest ratio between image quality and bit rate. Nevertheless, the Mobile and Calendar video sequence does not present the best quality-rate ratio for all the picture quality metrics.

One way to improve the performance of the Mux Bit algorithm regarding the vMean parameter would be to limit the bandwidth relocation among video programmes by establishing an additional threshold value.

In general, for the different joint coding algorithms, the value of the vMean parameter is positive for the IBBP GOP1 pattern and in most cases when the SSIM is used. These results suggest an average increase in the image quality perceived by the viewers. The following section presents the results of the subjective evaluation of image quality. Comparing simulation results between the two different GOP patterns, IBBP GOP1 presents better overall results than IPPP GOP1. The observation of the R-D behaviour of video sequences provides additional information to understand this. Consider, as an example, the Akiyo and Mobile and Calendar video sequences' performance when coded individually, at a constant bit rate (256kbps and 512 kbps) and with both GOP patterns. In the case of Akiyo, the image quality when the bit rate is reduced by half, from 512 kbps to 256 kbps, decreased 3.90 dB (PSNR) and 11.30 (SSIM) for IPPP GOP1 and 1.84 dB (PSNR) and 2.82 (SSIM) for IBBP GOP1. Thus, the same level of reduction in bit rate of Akiyo video sequence will generate a larger reduction for picture quality in sequences encoded with the IPPP GOP1 pattern compared with video sequences encoded with the IBBP GOP1 pattern. This reduction is even greater in the case of SSIM.

Another important issue is the nonlinearity of picture quality variation. For the Mobile and Calendar video sequence, SSIM increases 12.57 (IPPP GOP1) and 16.77 (IBBP GOP1) when bit rate increases from 256kbps to 512 kbps. Let us consider an intermediate value for the bit rate, 384kbps. When the bit rate changes from 256kbps to 384kbps, SSIM increases 5.85 (IPPP GOP1) and 10.51 (IBBP GOP1), and when the bit rate changes from 384 kbps to 512kbps, SSIM increases 6.72 (IPPP GOP1) and 6.26 (IBBP GOP1). Although in this last case the increase in SSIM is approximately the same value for both GOP patterns, when the bit rate changes from 256kbps to 384kbps, the variation in picture quality according to the SSIM metric is quite different with the two GOP patterns. Thus, it is very important to properly estimate a new R-D operation point. As a brief final remark, overall results support the use of SSIM as a viable metric in joint rate coding processes, particularly if using SSIM as the reference picture quality metric. Mux PSNR presents good results if PSNR. Mux PSPNR results are below SSIM results. Nevertheless, Mux PSPNR also achieves the goal of increasing the uniformization level of picture quality. PSPNR only takes into account the perceptible noise, that is, only distortion that exceeds the JND profile is taken into consideration ([187]). Changing the bit rate can affect

also the visibility of error in parts of the images. One way to improve Mux PSPNR performance would be to incorporate information such as the number of macroblocks that are in the boundary of visibility. Mux Bit presents good results for all the parameters when the complexity as defined by TM5 does not differ much among video sequences. When the complexity is quite heterogeneous, the performance of Mux Bit is characterised by the decrease of the differences in picture quality of the video sequences and a sharp increase of the minimum value. This is an interesting result if the goal is to maximize the minimum value of picture quality.

6.4 Subjective Video Quality Assessment

In the previous section, objective video quality assessment was used to evaluate the perceptual quality of video sequences. Objective assessment is a fast and cost-efficient way to evaluate visual quality, generating always the same output ([82]). Another advantage of objective assessment is that the result is independent of the context in which the quality of a video sequence is assessed. A single processed video sequence (PVS) can be evaluated by an objective quality metric, but it cannot be evaluated by a human viewer. A human viewer decides how to evaluate a video sequence by measuring it against a set of sequences. For example, a video sequence that can be assessed as ’excellent’ in an internet video streaming scenario, may be assessed as ’poor’ when assessed in a high-definition broadcast subjective test scenario. Objective assessment is very helpful for evaluating the progress in codec design of a specific algorithm.

At the same time, there are some problems as objective assessment is dependent on the type of codec used and the parameters defined. An objective assessment metric is not capable of providing the full truth about video codec quality ([82],[507]). Since the interest is in human opinions regarding video quality, subjective assessment is a reliable and a useful method of performing video quality assessment. However, subjective studies have to be conducted in a carefully controlled environment, and it is a time-consuming and expensive task because of the number of observers that need to be involved and the viewing conditions such as ambient illumination, display device, or viewing distance.

There are various methodologies that can be used by multimedia producers and television broadcasters to subjectively assess the quality of video programmes. By subjective video quality assessment, it is meant any methodologies in which several persons (or "observers") are involved in a process to assess video sequences under controlled conditions ([81],[507]). Note that subjective video quality assessment methodologies are different from the "expert viewing" methodologies. The "expert viewing" schemes are usually not suited to provide conclusive judgments on codec quality but rather for a fast evaluation or a first indicator of the level of

quality. Several methodologies were introduced in Chapter 2. A typical process of the subjective evaluation of the video codec performance can be resumed in the following phases ([81],[82],[507]):

- Select algorithms under evaluation (codec's with a particular set of encoding parameters such as pre-filtering, buffering, key frame distance, etc.).
- Select video test sequences (also frequently called SRC or Source Reference Channel or Circuit). Video sequences should vary motion and spatial detail characteristics and be available in uncompressed form.
- Define the settings of the system that is going to be evaluated (often called HRC or Hypothetical Reference Circuit). In broadcast applications, a HRC is a system composed of a video coder and a video decoder and sometimes includes a transmission distortion block. In the current case, it corresponds to defining the coding conditions such as bit rates and GOP patterns.
- Encode the video test sequences to generate coded representations of the test sequences.
- Select the assessment methodology and set up a reproducible environment (how sequences are display to viewers and how their opinion is gathered, for example: ACR, DSCQS, SSCQS, or SAMVIQ).
- Organize test sessions. Invite the panel of subjects (also called "observers," "assessors" or "evaluators"). The number of observers to invite should be large enough to guarantee that results are statistically relevant (depending on the assessment methodology, there could be a minimum of valid subject's participation). Carry out testing.
- Gather the assessment results, perform statistical analysis on gathered data and eliminate inconsistent subjects.
- Compute average marks for each HRC based on subjects' opinion.

Selecting the assessment methodology is a hard task. There are many ways of displaying video sequences to observers to gather their opinion and score. Some methods have been standardized and have been presented in Chapter 2. Two popular methods used to gather opinion scores are Absolute Category Rating (ACR) and Subjective Assessment Methodology for Video Quality (SAMVIQ). The ACR test method presents stimuli in a random order and uses a coarse resolution rating scale for evaluation. The SAMVIQ test method allows observers to view

several stimuli multiple times and uses a fine resolution rating scale for evaluation. Comparison studies between ACR and SAMVIQ have shown that ACR, and SAMVIQ provide correlated results for CIF sequences ([80]) and for an equal number of observers, SAMVIQ scores have greater accuracy than ACR scores (for an identical level of accuracy, on average, SAMVIQ required 30% fewer observers than ACR) ([508],[509]). Thus, SAMVIQ was selected for assessment.

6.4.1 *SAMVIQ Interface*

The underlying principle for the SAMVIQ methodology is explained by the considerable difference between the broadcast television and multimedia domains ([81],[82],[507]). Multimedia offers a wide-range of options over the rigid television domain architecture. While the television domain is quite specific, the multimedia domain consists of an array of options regarding codecs, image formats, frame rates and display types.

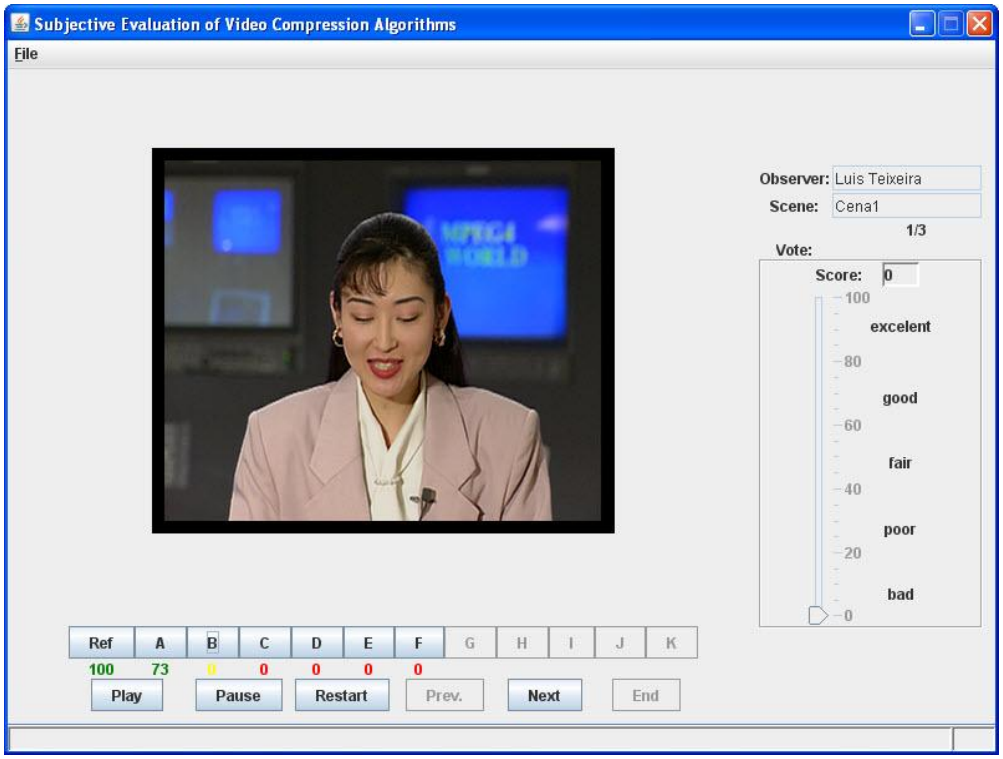


Figure 6.9 – SAMVIQ User Interface

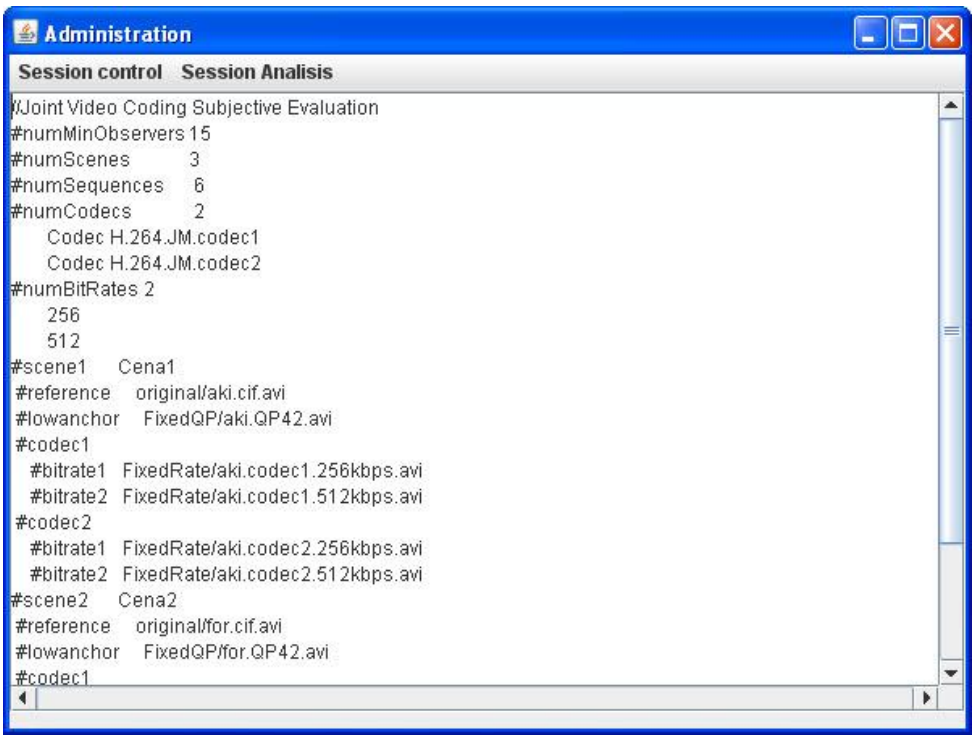


Figure 6.10 – SAMVIQ Administration Interface

These features associated with the possibility of multiple viewing distances have fostered the development of assessment methods dedicated to multimedia. SAMVIQ is a multi stimuli continuous quality scale method using explicit and hidden references ([81]). It generates a measure of the subjective video codec quality, which can be compared directly with the reference, i.e. the original video signal. Each test session is organized such that each participant evaluates scenes sequentially ([510]). For each scene under quality evaluation test, observers interact with the SAMVIQ graphical user interface for displaying and rating, any sequence in any order (Figure 6.9). The interface contains a series of command buttons that allow the user to play, pause and restart the viewing of a sequence, to assess a sequence and to select the next scene. On the left-hand bottom of the interface, there is a button, marked as Ref, that visibly identifies the explicit reference sequence (the original video signal). Buttons with letter labels (for example, A, B, C ...) give access to the remaining sequences for each scene. Except for the explicit reference sequence, the presentation order of each sequence in a scene is randomized. Through the graphical user interface, Figure 6.9, the observer can compare all coded versions of the scene as well as with the video reference, regulating the quality rating for each video sequence accordingly.

The observer can express their judgments by dragging a slider on a quality scale ranging from zero to one hundred. The highest quality should be marked "100" (top of the scale), and the lowest quality perceived should be marked "0" (bottom of the scale). The quality scale was

marked into sections. In addition, the quality scale was divided into five equal sections labelled with the adjectives: “Bad," “Poor," “Fair," “Good," and “Excellent." After the observers vote, the position of the slider was converted into a numeric mark by linearly mapping the scale to the interval [1, 100]. The observer can select the order by which videos are displayed and to alter the quality rating as many times as they want. Only the final opinion score is saved. Observers can only give a score after viewing the complete clip. The access to the next scene is obtained when the observer has successfully given his opinion score on all video clips at least once. The SAMVIQ interface was implemented in Java, and it consists of two applications: runAdmin (to specify the test organization and joint MOS results from different labs - Figure 6.10) and runObserver (the observer interface Figure 6.9).

6.4.2 *Test Organization*

Organization of test sessions is crucial for the success of the quality assessment experiments. In particular, the number and duration of each of the sessions for a given number of test sequences should be minimised. First, a set of reference video sequences must be selected. Source video sequences according to VQEG terminology, are referred to as Source Reference Channel or Circuits (SRC). Selected SRCs should vary in the amount of temporal information and spatial detail. These parameters affect the level of video compression of a sequence and as a result, the level of impairment that video sequences can suffer. When choosing SRCs, it can be helpful to compare the relative spatial information and temporal information of the different sequences. Usually, the compression difficulty is directly related to the spatial and temporal information of a sequence ([72]). One way to compare sequences is described in ITU-T Rec. P.910 which is to plot a xy-chart of the Spatial Information (SI) and Temporal Information (TI). When using a small set of video sequences, in a given test, it may be important to choose sequences that span a large part of the spatial-temporal information plane ([72]). On the other hand, if one were trying to select test sequences with similar coding difficulty, then sequences with similar SI and TI values should be selected. Figure 6.11 presents the spatio-temporal plot for the video test set sequences presented in previous Chapter. SI and TI were computed using formulas introduced in Chapter 2. Analysing the Temporal Information metric, video sequences can be divided into two groups: one containing the sequences Akiyo, Mother and Daughter, Hall, Silence, Paris e Deadline, and a second group with News, Mobile and Calendar, Coastguard, Foreman, Flower Garden and Football. Using spatial information, this last group can be divided into two sub-groups: News, Foreman and Football, with SI varying between 80 and 120, and a second sub-group composed of Coastguard, Mobile and Calendar and Flower Garden with SI varying

between 140 and 180. Akiyo presents the lowest values of TI and Football the highest value of TI. As for SI, Akiyo again presents the lowest value and Flower Garden the highest value.

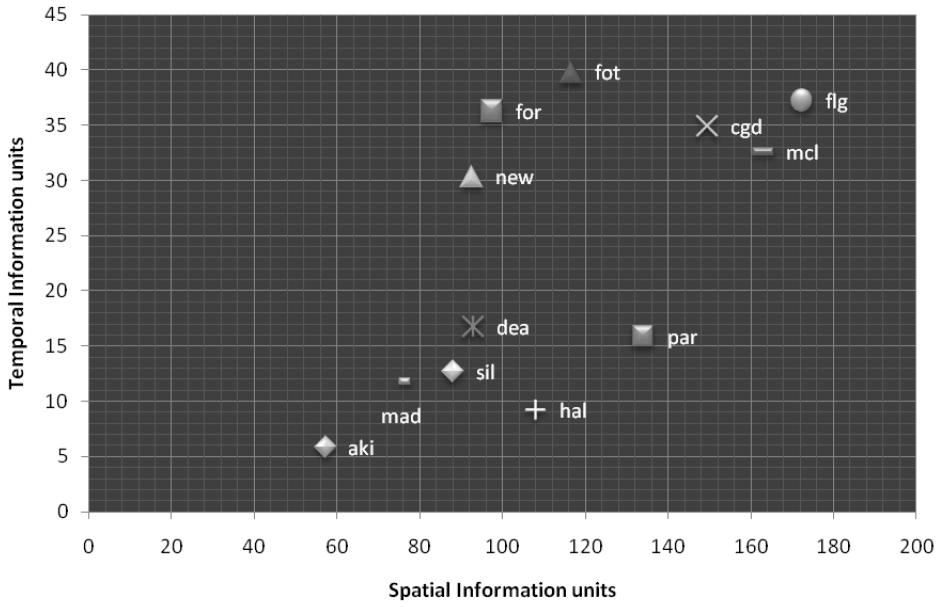


Figure 6.11 – Spatial-temporal plot for video test sequence set

After selecting the video test sequences, it is necessary to define the settings of the system to be assessed. VQEG suggests using a matrix of SRC×HRC, so that each SRC is processed by each HRC.

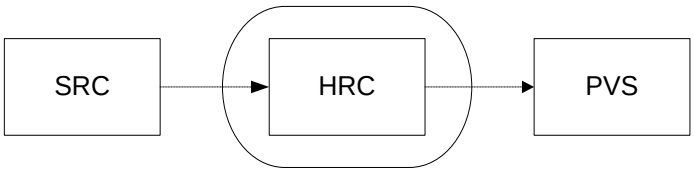


Figure 6.12 – SRSs, HRCs, and PVSs

The application of a HRC to a SRC generates a Processed Video Sequence (PVS) (Figure 6.12). A PVS is obtained by encoding a reference video through a specific coding setup (in the current case the joint coding algorithm, reference bit rate and GOP Pattern). Each SAMVIQ session consists of PVSs and explicit reference (high anchor), hidden reference (high anchor) and low anchor. High and low anchors are used to show observers what are the limits of the quality scale ([81]). The hidden and explicit references are identical and are the uncompressed version of the sequence. The explicit reference is visibly labelled as the reference for the observer, while the other sequences are not labelled. The low anchor corresponds to a low-end codec. Quality anchors are included to stabilize the results ([510]). Explicit reference to minimised standard deviations of scores, and the hidden reference assesses the intrinsic quality of the reference.

Without an explicit reference, the standard deviation would be dramatically increased ([510]). A test session organization is as follows (Figure 6.13) ([82]).

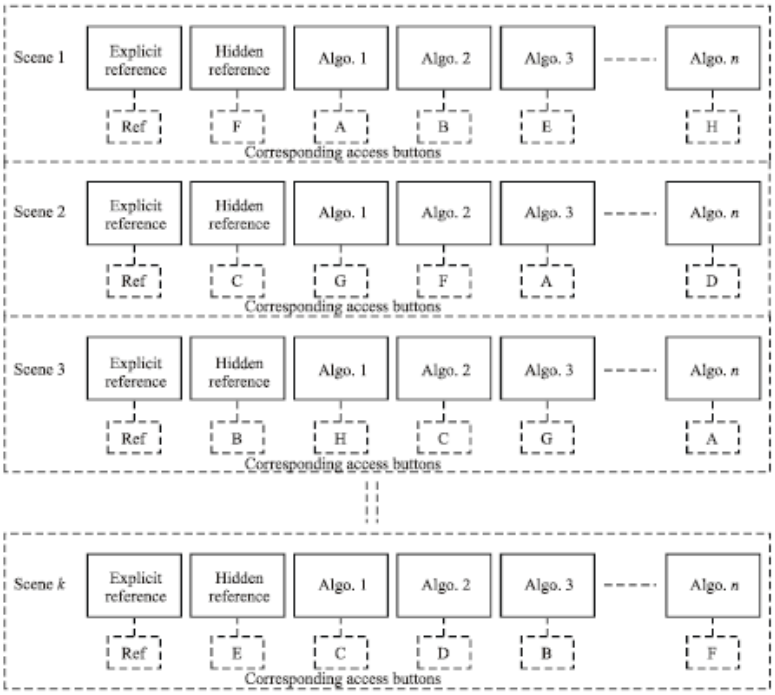


Figure 6.13 – Test organization example for SAMVIQ method ([510])

The number of SRCs and HRCs is limited by the duration of the subjective test. SAMVIQ has a limit of 10 HCR for each scene, excluding the explicit and hidden references ([81],[507]). SAMVIQ also specifies that the limit duration of a session is 30 minutes due to the drop of concentration of observers ([81]). Nevertheless, SAMVIQ allows an unlimited number of views. Quan reports in [507] that as observers may want to view some of the SRCs more than one time, for a SRC with eight seconds duration, it is not possible to run much more than 90 PVS in any single SAMVIQ test session. For SRCs with 10 seconds duration, the equivalent maximum number of PVS is 72 PVS per session. In previous sections, the simulation results from the joint video coding of two, three and six video sources were analysed. For each scenario, several parameters were altered: video sources, reference bit rates (256kbps and 512kbps), GOP patterns (four different GOP patterns), and five different encoding methods (independent coding, Mux Bit, Mux PSNR, Mux SSIM and Mux PSPNR). As a result, 240 PVS were generated in the first joint coding scenario (two video programmes), 600 PVS in the second joint coding scenario (three video programmes), and 120 PVS in the third joint coding scenario (six video programmes). Considering that a session should not exceed 72 PVS, then to assess all the simulation results 14 sessions and a minimum of 210 different observers would be needed. SAMVIQ suggests that in order to obtain stable results, the minimum number of scenes should be four ([81],[507]). It was decided to perform the subjective test only on the third joint coding

scenario. Therefore, the number of sessions was reduced from 14 to 2 sessions, and the minimum number of different observers from 210 to 30. In the first session, all the PVS were encoded with IBBP GOP1 and in second session with IPPP GOP1. 10 HRC per SRC were selected, corresponding to the five coding algorithms and the two reference bit rates (256kbps and 512kbps).

HRC nr	Codec	Bit rate	Frame Rate	Spatial Resolution	GOP Pattern	other
0	None	none	30 fps	CIF	none	reference
1	H.264.JM11	256 kbps	30 fps	CIF	IBBP GOP1	
2	H.264.JM11	512 kbps	30 fps	CIF	IBBP GOP1	
3	H.264.Mux Bit	256 kbps	30 fps	CIF	IBBP GOP1	
4	H.264.Mux Bit	512 kbps	30 fps	CIF	IBBP GOP1	
5	H.264.Mux PSNR	256 kbps	30 fps	CIF	IBBP GOP1	
6	H.264.Mux PSNR	512 kbps	30 fps	CIF	IBBP GOP1	
7	H.264.Mux SSIM	256 kbps	30 fps	CIF	IBBP GOP1	
8	H.264.Mux SSIM	512 kbps	30 fps	CIF	IBBP GOP1	
9	H.264.Mux PSPNR	256 kbps	30 fps	CIF	IBBP GOP1	
10	H.264.Mux PSPNR	512 kbps	30 fps	CIF	IBBP GOP1	
11	H.264.QP42	--	30 fps	CIF	IBBP GOP1	low anchor

Table 6.28 – List of HCR for IBBP GOP1

Table 6.28 shows the different HRCs for the first session (IBBP GOP1). In the second session the GOP pattern was altered from IBBP GOP1 to IPPP GOP1. HRC0 refers to the hidden reference condition in the SAMVIQ method and HRC11 to the low anchor. In this case, the low anchor corresponds to encoding SRC with a fixed quantisation parameter (QP = 42). All video sequences were converted and stored in uncompressed AVI- RGB24 format.

SRC nr	SRC1	SRC2	SRC3	SRC4	SRC5	SRC6
Video Sequence Name	Akiyo	Football	Hall	MAD	MCL	Silence

Table 6.29 – List of SRC

Table 6.29 shows the list of SRC used in the first and second sessions. Testing took place using a Microsoft Windows computer. The video sequences were displayed on a LCD monitor (HP L2045w). Table 6.30 summarizes the viewing conditions and the monitor specifications.

Parameter	Setting
Display technology	LCD display / TFT active matrix
Reference name of the display	RB145AA - HP L2045w
Viewing distance	Not constrained: front of display
Image Brightness	300 cd/m²
Image Colour Temperature	6500K
Image Contrast ratio	800:1
Display size (diagonal in inches)	20.1” in
Screen Resolution	1024x768
Refresh rate	60 Hz
Dot Pitch / Pixel Pitch	0.258 mm
Response time	5 ms

Table 6.30 – Viewing conditions and monitor specifications

The computer was in an office environment with normal indoor illumination levels. Viewing distance was not fixed, as SAMVIQ does not require any specific viewing distance range ([81]). Each observer was allowed to adjust his own optimal viewing distance according to his preference for comfortable viewing. Nevertheless, during the training phase, subjects were asked to maintain their back in contact with the chair. The chair was initially positioned at a distance of 4H from the screen (where H is the picture height). In total, 30 observers, aged 18-22 participated in the tests. Observers were university students from the Portuguese Catholic University, School of Arts, with experience in using a computer and with Smartphones and mobile phones with a video camera. No expert observers participated in the evaluation. Often, expert observers have preconceived judgements of video artefacts, resulting in somewhat biased scoring ([82]). None of the subjects had taken part in any subjective testing or had previous experience in fields related to video coding or assessing picture quality.

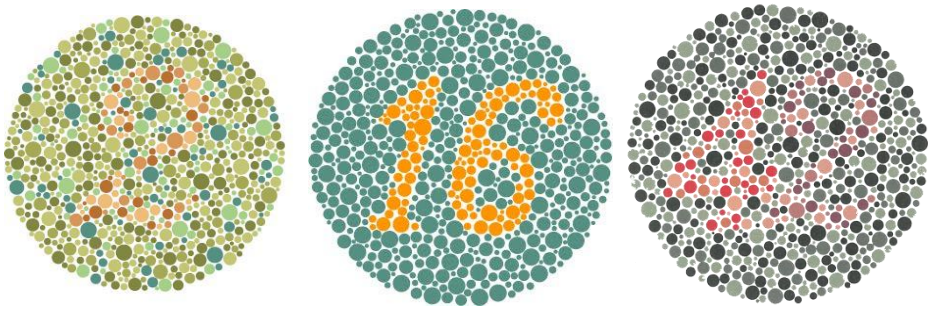


Figure 6.14 – Ishihara colour plates

In a subjective assessment session, observers should be evaluated for their visual acuity. Before the session began, the observers were screened for normal visual acuity on the Snellen chart and for normal colour vision using Ishihara colour plates (Figure 6.14). All observers were reported to have a normal vision.

The subjective assessment session started with a brief introduction, reading out the instructions, to explain the goals of the test and a short demonstration to present the voting method ([72],[81]).

"Welcome. In this session, you will be asked to assess the overall quality of a set of short video sequences by inserting a slider mark on a vertical scale. The grading scale is continuous and is divided in five equal parts, as follows: Excellent (80 to 100 points), Good (60 to 80 points), Fair (40 to 60 points), Poor (20 to 40 points) and Bad (0 to 20 points). Observe carefully the entire video sequence before making your judgment. Please do not support your judgment based on the content of the scene. Keep in mind the various aspects of the video quality and develop your opinion based upon your total idea of the video quality. You can only assess a video sequence after having viewed the entire video sequence. You can only proceed to the next sequence after having assessed all video clips."

Both ITU-R BT.500 and EBU SAMVIQ recommends that observers should be carefully initiated into the method of assessment, the types of impairment or quality factors likely to take place, the grading scale, timing, etc. One way to minimise the learning effects is to include a few "dummy presentations" at the beginning of each test session. These training sequences should demonstrate the range and the type of the impairments that the observers are going to assess during the session. Training presentations must not be including in the statistical analysis of test results.

6.4.3 *Statistical Analysis*

Depending on the selected assessment method, the first step before performing the statistical analysis of results is to determine the normalised opinion scores. The normalization process is based on the opinion score of the video references. VQEG's multimedia group in their multimedia tests plan ([511]) defines a normalization process for each observer and for each PVS.

$$V'_{a,s,h} = \begin{cases} 1 & if\ V_{a,s,ref} > V_{a,s,h} \\ \dfrac{V_{a,s,h}}{V_{a,s,ref}} & otherwise \end{cases} \tag{6.53}$$

Equation (6.53) presents the method to normalize the opinion scores, where $V_{a,s,h}$ designates the opinion score given by the observer a to the $PVS_{s,h}$ (corresponding to SRC s and HRC h); $V_{a,s,ref}$ is the opinion score given by the same observer a to the reference video of the SRC s; and $V'_{a,s,h}$ is the normalised opinion score of the observer regarding $PVS_{s,h}$. The normalised

process should be applied for all PVSs and all observers prior to any other analysis. After the normalization process, opinion scores will be in the interval $0 \le V'_{a,s,h} \le 1$. Another normalization process is DMOS (Differential Mean Opinion Score) that is reported to present similar results ([72],[511]). Differential observer opinion scores are calculated on a per subject per processed video sequence basis. The corresponding hidden reference is used to calculate DMOS using the following formula where K corresponds to the scale.

$$V'_{a,s,h} = V_{a,s,h} - V_{a,s,ref} + K \tag{6.54}$$

After the normalization and before the statistical analysis, all the participants in the subjective assessment session must be examined in order to determine the consistency of their scores; this is, to guarantee that the observer did not vote randomly. This occasionally happens because the objective of the assessment was not properly explained or understood. SAMVIQ specifies a rejection criterion that verifies the level of consistency of the opinion scores of one observer according to the mean opinion score (MOS) of all observers for a given test session ([81]). If the relationship between the quality scale and score range of observers is supposed to be linear, then SAMVIQ proposes to use Pearson's correlation (Equation (6.55))

$$r(x,y) = \frac{\left(\sum_{i=1}^{n} x_i y_i\right) - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sqrt{\left(\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right)\left(\sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}\right)}} \tag{6.55}$$

where x_i is the mean score of all observers for the triplet (algorithm-bit rate-scene), y_i the individual score of one observer for the same triplet, n the product between the number of algorithms and the number of scenes, and i corresponds to the index {codec number, bit rate number, scene number} ([81]). The hidden reference can be considered as a high quality anchor. If the low and high anchors are included, they increase the correlation score, conversely the correlation offsets between the observers are decreased ([81]). If the relationship between the quality scale and score range of observers is not supposed to be linear than Spearman rank correlation should be used as follows:

$$r(x,y) = \left[1 - \frac{6 \times \sum_{i=1}^{n} \left[R(x_i) - R(y_i) \right]^2}{n^3 - n} \right] \tag{6.56}$$

where $R(x_i$ or $y_i)$ corresponds to the ranking order. Final rejection criteria for discarding an observer from a test is defined in Equation (6.57) and Equation (6.58).

$$\begin{aligned} &\text{IF } \left[\text{mean}(\text{r}) - \text{std}(\text{r}) \right] > \text{MCT} \\ &\text{THEN Rejection threshold} = \text{MCT} \\ &\text{ELSE Rejection threshold} = \left[\text{mean}(\text{r}) - \text{std}(\text{r}) \right] \end{aligned} \tag{6.57}$$

$$\begin{aligned} &\text{IF } \left[\text{r} \left(\text{Observer}_i \right) \right] > \text{Rejection threshold} \\ &\text{THEN observer “i” of the test is not discarded} \\ &\text{ELSE observer “i” of the test is discarded} \end{aligned} \tag{6.58}$$

where:

MCT (Max Correlation Threshold) = 0.85

$\text{r} = \min \left(\text{Pearson correlation}, \text{Spearman rank correlation} \right)$

$\text{mean}(\text{r})$ is the average of the correlations of all the observers of a test

$\text{std}(\text{r})$ is the standard deviation of all observers' correlations of a test

As SAMVIQ does not specify the use of a normalization process, it was decided to perform the Pearson correlation analysis both on raw data and on normalized MOS using equation (6.53). The average of all observer's opinion scores was determined for each PVS and for all PVSs of a single HRC, resulting in the mean value of SRCs of each HRC. For each observer, the correlation to the mean of all observers was determined for SRC and HRC distributions.

	a1	a2	a3	a4	a5	a6	a7	a8	a9	a10	a11	a12	a13	a14	a15
HRC	0.96	0.93	0.88	0.90	0.87	0.95	0.90	0.92	0.86	0.88	0.86	0.87	0.90	0.91	0.86
SRC	0.97	0.92	0.95	0.95	0.98	0.96	0.99	1.00	0.97	0.97	0.90	0.99	0.93	0.94	0.96

Table 6.31 – Pearson Correlation Analyses per Observer (IBBP GOP1)

	a1	a2	a3	a4	a5	a6	a7	a8	a9	a10	a11	a12	a13	a14	a15
HRC	0.96	0.95	0.93	0.94	0.98	0.97	0.95	0.95	0.94	0.97	0.98	0.97	0.95	0.98	0.93
SRC	0.95	0.92	0.93	0.98	0.98	0.99	0.99	0.99	0.98	0.98	0.93	0.99	0.93	0.94	0.96

Table 6.32 – Pearson Correlation Analyses per Observer (IPPP GOP1)

Table 6.31 and Table 6.32 show respectively the correlation between each observer's individual opinion scores and the mean value of all other observers' opinion scores, both per HRC and per

SRC for the first and second sessions. Results for all the observers are above the SAMVIQ correlation threshold limit (0.85). No observer's opinion scores were discarded.

	a1	a2	a3	a4	a5	a6	a7	a8	a9	a10	a11	a12	a13	a14	a15
HRC	0.96	0.92	0.89	0.90	0.87	0.95	0.90	0.92	0.85	0.87	0.85	0.87	0.90	0.91	0.86
SRC	0.94	0.92	0.96	0.89	0.98	0.98	0.98	0.99	0.92	0.94	0.86	0.99	0.95	0.94	0.97

Table 6.33 – Pearson Correlation Analysis per Observer (IBBP GOP1, normalised opinion score)

	a1	a2	a3	a4	a5	a6	a7	a8	a9	a10	a11	a12	a13	a14	a15
HRC	0.96	0.95	0.93	0.94	0.98	0.97	0.95	0.95	0.94	0.97	0.98	0.97	0.95	0.98	0.93
SRC	0.93	0.92	0.91	0.98	0.98	0.99	0.99	0.99	0.96	0.98	0.93	0.97	0.92	0.94	0.97

Table 6.34 – Pearson Correlation Analysis per Observer (IPPP GOP1, normalised opinion score)

Normalised opinion score per observer regarding each SRC and HRC are shown in Table 6.33 and Table 6.34. Comparing Pearson correlation analysis results applied to raw data with results applied to normalised opinion score, it is observed that the differences found are rather small, with both methods generating similar results. So it was decided to present in this section normalised opinion score results and in Annex C, the non-normalised results.

After discarding the outliers, the next step is to analyse results using the MOS measure. ITU-R Rec. BT.500 defines MOS as the mean of the observers' opinion scores ([71]). The MOS can be calculated for each PVS, each HRC or each SRC. The MOS of a PVS, MOS_{sh}, is the mean of the opinion scores attributed by all observers to a particular PVS (corresponding to HRC h and SRC s). It can be determined as follows:

$$MOS_p = MOS_{sh} = \frac{1}{N}\sum_{a=1}^{N} V'_{a,s,h} \tag{6.59}$$

where $V'_{a,s,h}$ is the normalised opinion score of an observer a, for a given test condition h (HCR=h), and video sequence s (SRC = s) and N is the number of observers. In the same way, overall mean opinion scores, MOS_s and MOS_h, are computed for each test video sequence and for each test condition. The MOS_s shows the mean quality of each source video, while the MOS_h is important to assess the mean quality perceived in each variation applied to the video sources. SAMVIQ specifies that all mean opinion scores should have an associated confidence interval (CI). In the case of MOS_{sh}, δ_{sh} is the associated confidence interval so results should be presented as follows:

$$\left[MOS_{sh} - \delta_{sh}, MOS_{sh} + \delta_{sh}\right] \tag{6.60}$$

SAMVIQ proposes 95% confidence intervals given by

$$\delta_{sh} = t_{0.05} \cdot \frac{std_{sh}}{\sqrt{N}} \tag{6.61}$$

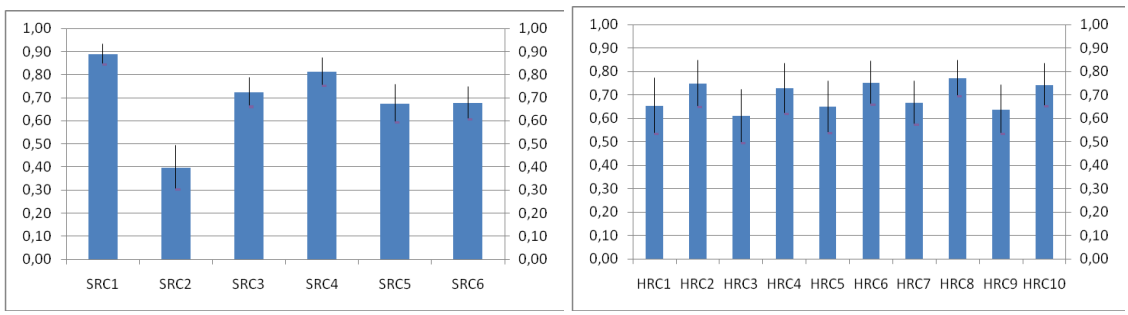
where $t_{0.05}$ is the t-value associated for the desired significance level of 95% with N − 1 degrees of freedom, and std_{sh} is the standard deviation of each presentation. Table 6.35 and Table 6.36 present results for all SRC per HRC for both sessions (IPPP GOP1 and IBBP GOP1). MOS results per observer for SRC and for HRC are also available in Annex C. Figure 6.16 shows a plot of normalised MOS obtained for each of the SRCs of both assessment sessions (IBBP GOP1 and IPPP GOP1) and for each of the HRC. Blue bars represent the MOS and the vertical blue segments the associated 95% confidence interval.

	IBBP GOP1				IPPP GOP1			
HRC	μ	CI	σ	CoV	μ	CI	σ	CoV
SRC1	0.888	0.045	0.089	0.100	0.736	0.057	0.114	0.155
SRC2	0.397	0.096	0.189	0.476	0.325	0.100	0.197	0.606
SRC3	0.724	0.064	0.127	0.175	0.569	0.072	0.141	0.248
SRC4	0.812	0.060	0.119	0.147	0.690	0.072	0.142	0.206
SRC5	0.675	0.083	0.164	0.243	0.278	0.092	0.183	0.658
SRC6	0.676	0.071	0.140	0.207	0.479	0.097	0.192	0.401

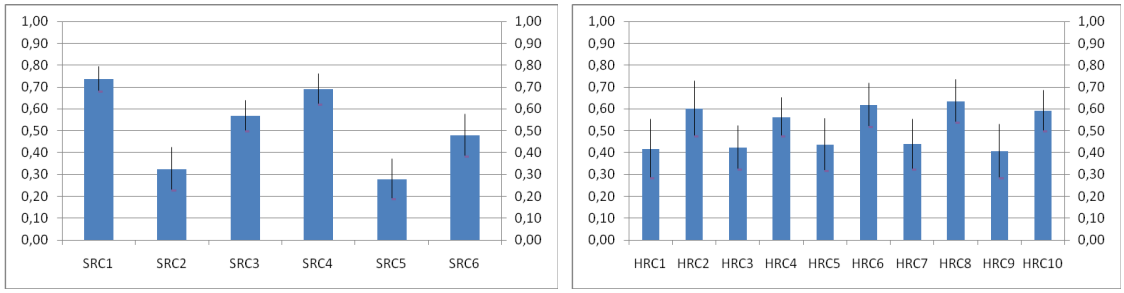
Table 6.35 – Mean (μ), Confidence Interval at 95% (CI) and Standard Deviation (σ) for MOS_s (IBBP GOP 1 and IPPP GOP1)

	IBBP GOP1				IPPP GOP1			
HRC	μ	CI	σ	CoV	μ	CI	σ	CoV
HRC1	0.653	0.121	0.239	0.366	0.418	0.135	0.266	0.636
HRC2	0.748	0.101	0.200	0.267	0.601	0.128	0.252	0.419
HRC3	0.610	0.115	0.227	0.372	0.423	0.102	0.201	0.475
HRC4	0.727	0.107	0.212	0.292	0.563	0.088	0.174	0.309
HRC5	0.649	0.111	0.219	0.337	0.436	0.120	0.238	0.546
HRC6	0.751	0.093	0.185	0.246	0.617	0.100	0.197	0.319
HRC7	0.665	0.094	0.186	0.280	0.438	0.116	0.229	0.523
HRC8	0.771	0.077	0.153	0.198	0.634	0.099	0.195	0.308
HRC9	0.637	0.105	0.208	0.327	0.407	0.124	0.244	0.600
HRC10	0.743	0.092	0.182	0.245	0.592	0.094	0.186	0.314

Table 6.36 – Mean (μ), Confidence Interval at 95% (CI) and Standard Deviation (σ) for MOS_h (IBBP GOP1 and IPPP GOP1)



a) MOS_s (IBBP GOP1) b) MOS_h (IBBP GOP1)



c) MOS_s (IPPP GOP1) d) MOS_h (IPPP GOP1)

Figure 6.15 – MOS_s and MOS_h values with 95% CI (IBBP GOP1, IPPP GOP1)

Analyzing the MOS_s results (Table 6.35) it can be seen that the IPPP GOP1 SRCs present lower mean scores and higher standard deviation values compared to IBBP GOP1 SRCs. The mean picture quality and the mean standard deviation for the PVSs processed using the IBBP GOP1 pattern is 0.695 and 0.138 respectively, while for the IPPP GOP1 pattern the values are correspondingly 0.513 and 0.162. Furthermore, for IBBP GOP1 the higher the picture quality of the SRC, the lower is the value of the standard deviation. This type of relationship is not observed for IPPP GOP1. Thus, to allow a better comparison of the "relative variability," it was decided to compute CoV for both cases, IPPP GOP1 and IBBP GOP1. In both cases, the lower the mean picture quality of the SRC, the higher is the CoV. Best results regarding picture quality are obtained with sequences of lower complexity, such as Akiyo (SRC1). On the other extreme, the worst results are observed in more spatio-temporally complex sequences such as Football (SRC2) or MCL (SRC5). This can be interpreted as a difficulty of the observers to assess the picture quality of the SRCs with inferior picture quality or a stronger disagreement of opinions among observers when picture quality decreases. These SRCs present higher spatio-temporal complex characteristics.

Analyzing the MOS_h results (Table 6.36) corresponds to examine the different joint coding strategies (Table 6.28). Results show, in both GOP patterns, that the mean values of standard deviation and CoV are lower for video sequences encoded at 512kbps than for video sequences encoded at 256kbps. Thus, for higher bit rates, observers seem to agree more easily. On average, Mux SSIM and MUX PSNR present the best subjective picture quality results. Furthermore, the results of subjective image quality of the Mux PSNR algorithm and the Mux SSIM algorithm exceed, on average, the values of subjective image quality for the H.264 video encoder, using JM rate control, working at constant bit rate, in both the subjective assessment sessions. Typically, the Mux Bit algorithm presents, on average, the lowest results of subjective image quality, regardless of the GOP pattern or bit rate.

As for the IBBP GOP1 pattern, the standard deviation decreases when the reference bit rate increases for each joint coding algorithm. The Mux SSIM algorithm presents the highest average values of subjective quality of image, and the lowest values of standard deviation and CI. The joint coding algorithms can be sorted in decreasing order according to the subjective image quality scores, when the reference bit rate is 512 kbps, as follows: Mux SSIM, Mux PSNR, H.264 JM, Mux PSPNR, and Mux Bit. For 256kbps, the order is as follows: Mux SSIM, H.264 JM, Mux PSNR, Mux PSPNR, and Mux Bit. The interval of variation of the average values of the subjective image quality scores, regarding the IBBP GOP1 pattern, varies from 0.727 to 0.771 (512kbps), and from 0.610 to 0.665 (256kbps).

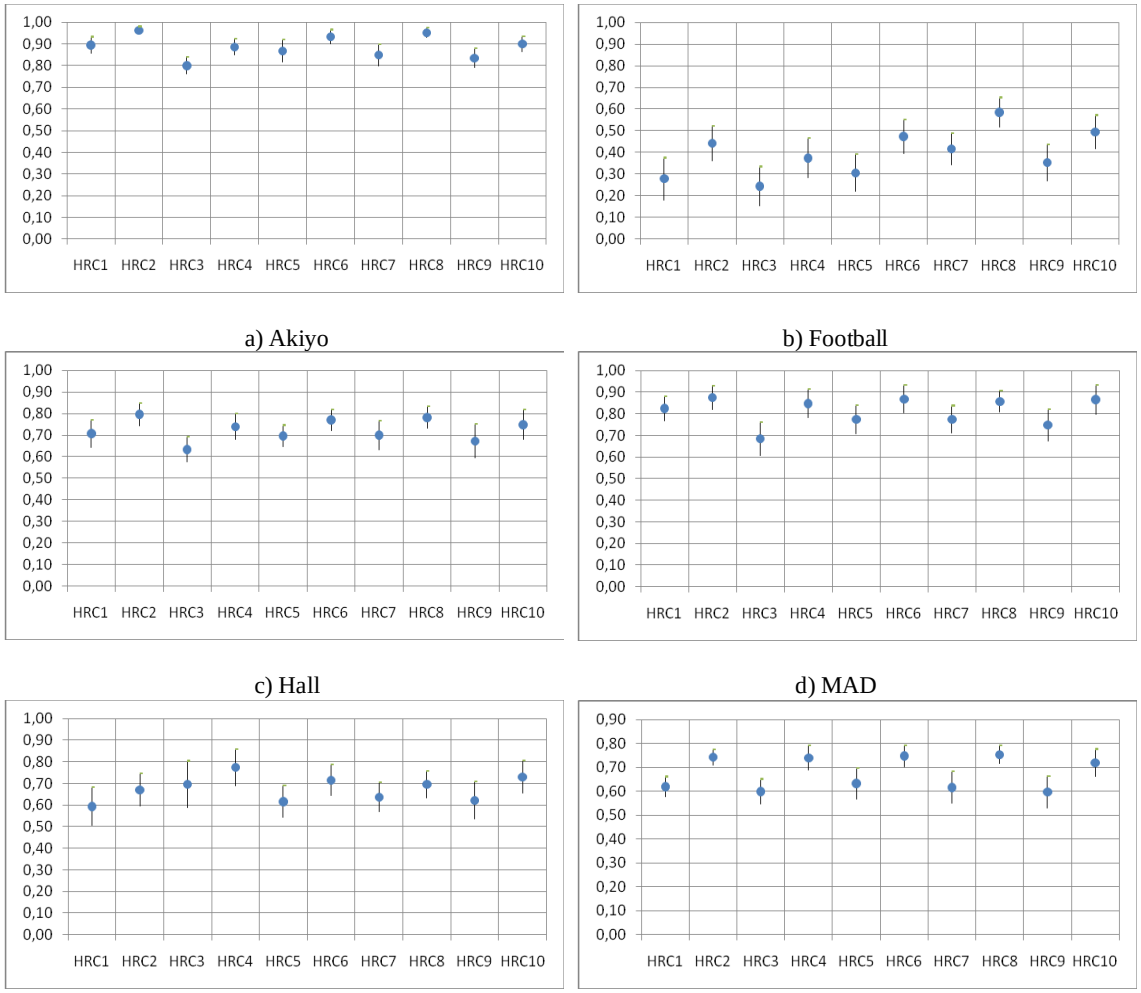
It can be seen for the IPPP GOP1 pattern that the HRCs have the lower mean scores and scores are further from the mean score, when compared with IBBP GOP1 results. Thus, the mean standard deviation for the PVSs processed using the IPPP GOP1 pattern is higher. This can be understood as a difficulty of the viewers in evaluating the quality of the videos or different opinions about them. Otherwise, the HRCs with higher scores (IBBP GOP1) are generally closer to the mean. These statements show that viewers have more agreement on the quality assessment of video's sequences when the value of the quality is higher than video sequences with lower values of the quality. The Mux SSIM algorithm displays again the highest average values of the subjective quality of image, and lower values of the standard deviation and CI parameters. The joint coding algorithms can be sorted in decreasing order according to the subjective image quality scores, when the reference bit rate is 512 kbps, as follows: Mux SSIM, Mux PSNR, H.264 JM, Mux PSPNR, and Mux Bit (the same order as IBBP GOP1). As for 256kbps, the order is as follows: Mux SSIM, Mux PSNR, Mux Bit, H.264 JM, and Mux PSPNR (the order differs from IBBP GOP1). The interval of variation of the average values of the subjective image quality scores, regarding the IBBP GOP1 pattern, varies from 0.563 to 0.634 (512kbps), and from 0.407 to 0.438 (256kbps).

Finally, results are analyzed for the different video sources, according to GOP pattern and bit rate. The normalised MOS values and their confidence intervals, at 95%, for each SRC, are shown in Figure 6.16 (IBBP GOP1) and Figure 6.17 (IPPP GOP1). In each plot, the centre of the bars indicates the value of the MOS observation, and the blue line segments represent the confidence interval around it. The numerical values of the mean (μ), confidence interval 95% (CI) and standard deviation (σ) for PVSs are available in Table 6.37 (IBBP GOP1) and Table 6.38 (IPPP GOP1).

In the two testing sessions, each video sequence was encoded with the H.264 JM algorithm (HRC1 and HRC2). Thus, it is possible to identify when joint coding algorithms present higher score compared to independent coding.

Regarding IBBP GOP1 (512kbps), three video sources present better subjective scores than when video sources are independently encoded: SRC2 (fot), SRC5 (mcl) and SRC6 (sil). In detail, SRC5 improves its performance for HRC4 (Mux Bit - 512kbps), HRC6 (Mux PSNR - 512kbps), HRC8 (Mux SSIM - 512kbps), and HRC10 (Mux PSPNR - 512kbps); SRC2 obtains better results for HRC6, HRC8 and HRC10; and SRC6 improves its performance for HRC6 and HRC8. As for IBBP GOP1 (256kbps), the results are quite similar regarding the joint coding algorithms. The main difference occurs in SRC6. While at 512kbps, scores improve with Mux PSNR and Mux SSIM, for 256kbps, scores only improve with Mux PSNR.

Besides the numeric score given by the observers, the quality scale was also divided into five equal intervals with the following adjectives from top to bottom: "Excellent" 100-80, "Good" 79-60, "Fair" 59-40, "Poor" 39-20 and "Bad" 19-0.



a) Akiyo b) Football

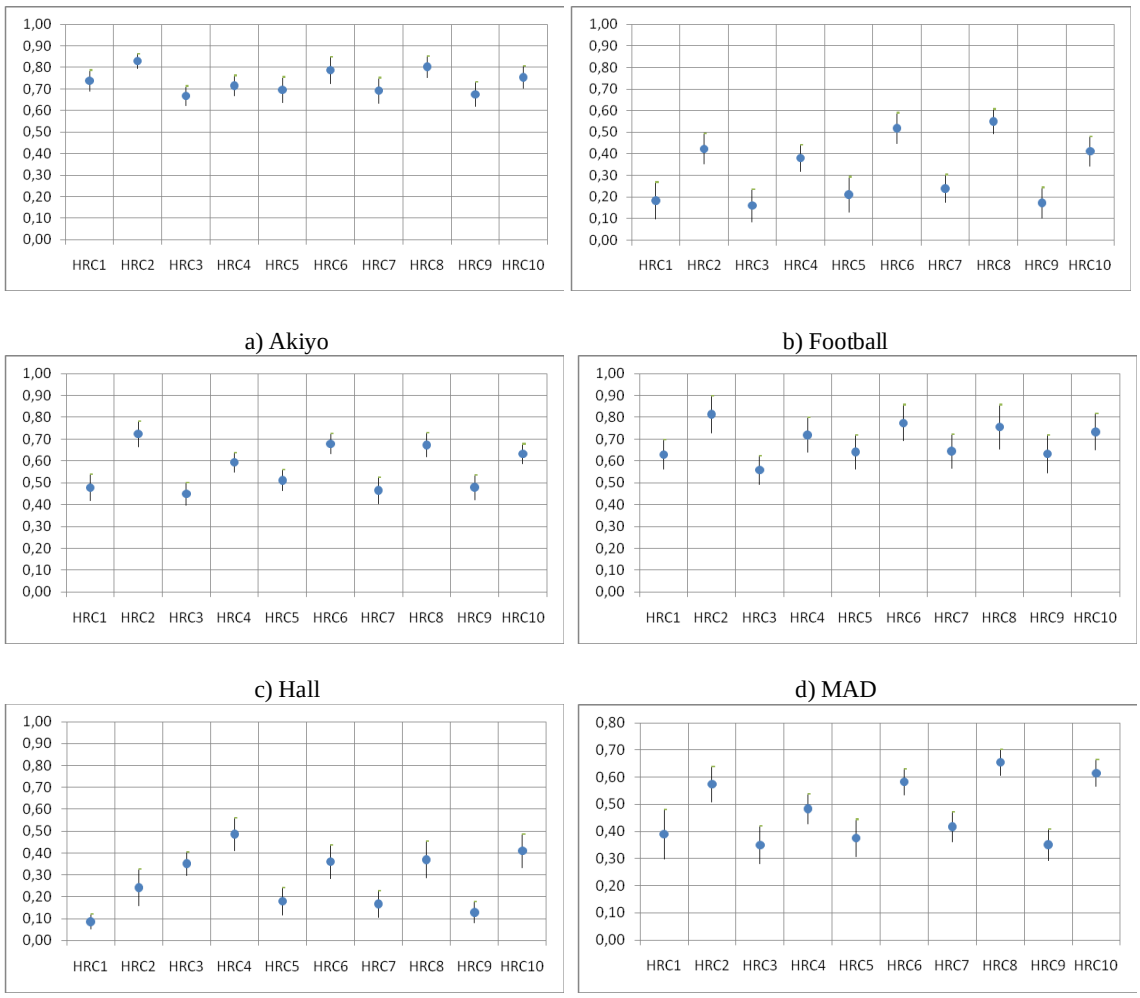
c) Hall d) MAD

e) MCL f) SIL

Figure 6.16 – Normalised MOS values and 95% CI for SRC Akiyo (a), Football (b), Hall (c), Mother and Daughter (d), Mobile and Calendar (e) and Silence (f) (IBBP GOP1)

SRC	SRC1 (Akiyo)			SRC2 (Fot)			SRC3 (Hall)			SRC4 (MAD)			SRC5 (MCL)			SRC6 (SIL)		
HRC	μ	CI	σ	μ	CI	σ	μ	CI	σ	μ	CI	σ	μ	CI	σ	μ	CI	σ
HRC1	0.90	0.04	0.07	0.28	0.10	0.19	0.71	0.06	0.12	0.82	0.04	0.08	0.59	0.09	0.18	0.62	0.06	0.11
HRC2	0.96	0.02	0.04	0.44	0.08	0.16	0.80	0.05	0.10	0.88	0.03	0.07	0.67	0.08	0.15	0.74	0.05	0.11
HRC3	0.80	0.04	0.08	0.24	0.09	0.18	0.63	0.06	0.11	0.68	0.05	0.10	0.70	0.11	0.21	0.60	0.08	0.15
HRC4	0.89	0.04	0.07	0.37	0.09	0.18	0.74	0.06	0.12	0.85	0.05	0.10	0.77	0.09	0.17	0.74	0.07	0.13
HRC5	0.87	0.05	0.10	0.31	0.09	0.17	0.70	0.05	0.10	0.77	0.07	0.13	0.62	0.07	0.15	0.63	0.07	0.13
HRC6	0.93	0.03	0.07	0.47	0.08	0.15	0.77	0.05	0.10	0.87	0.04	0.09	0.72	0.07	0.14	0.75	0.06	0.13
HRC7	0.85	0.05	0.10	0.42	0.07	0.15	0.70	0.07	0.14	0.77	0.07	0.13	0.64	0.07	0.13	0.62	0.06	0.12
HRC8	0.95	0.02	0.05	0.58	0.07	0.14	0.78	0.05	0.10	0.86	0.04	0.08	0.69	0.06	0.12	0.75	0.05	0.10
HRC9	0.83	0.05	0.09	0.35	0.08	0.17	0.67	0.08	0.16	0.75	0.07	0.13	0.62	0.09	0.17	0.60	0.07	0.15
HRC10	0.90	0.03	0.07	0.49	0.08	0.15	0.75	0.07	0.14	0.87	0.06	0.12	0.73	0.08	0.15	0.72	0.07	0.14

Table 6.37 – Mean (μ), Confidence of Interval at 95% (CI) and Standard Deviation (σ) of all SRCs per HRC (IBBP GOP1)



a) Akiyo b) Football

c) Hall d) MAD

e) MCL f) SIL

Figure 6.17 – Normalised MOS values and 95% CI for SRC Akiyo (a), Football (b), Hall (c), Mother and Daughter (d), Mobile and Calendar (e) and Silence (f) (IPPP GOP1)

SRC	SRC1 (Akiyo)			SRC2 (Fot)			SRC3 (Hall)			SRC4 (MAD)			SRC5 (MCL)			SRC6 (SIL)		
HRC	μ	CI	σ	μ	CI	σ	μ	CI	σ	μ	CI	σ	μ	CI	σ	μ	CI	σ
HRC1	0.74	0.05	0.10	0.18	0.09	0.17	0.48	0.06	0.12	0.63	0.09	0.18	0.09	0.03	0.07	0.39	0.07	0.14
HRC2	0.83	0.04	0.07	0.42	0.07	0.14	0.72	0.06	0.11	0.81	0.07	0.13	0.24	0.08	0.17	0.57	0.08	0.17
HRC3	0.67	0.05	0.09	0.16	0.08	0.15	0.45	0.05	0.10	0.56	0.07	0.14	0.35	0.05	0.11	0.35	0.06	0.13
HRC4	0.72	0.05	0.09	0.38	0.06	0.12	0.59	0.04	0.09	0.72	0.06	0.11	0.49	0.08	0.15	0.48	0.08	0.16
HRC5	0.70	0.06	0.12	0.21	0.08	0.16	0.51	0.05	0.09	0.64	0.07	0.14	0.18	0.06	0.13	0.38	0.08	0.15
HRC6	0.79	0.06	0.12	0.52	0.07	0.14	0.68	0.05	0.09	0.77	0.05	0.09	0.36	0.08	0.15	0.58	0.08	0.17
HRC7	0.69	0.06	0.12	0.24	0.06	0.13	0.47	0.06	0.12	0.65	0.06	0.11	0.17	0.06	0.12	0.42	0.08	0.15
HRC8	0.80	0.05	0.10	0.55	0.06	0.11	0.67	0.06	0.11	0.76	0.05	0.09	0.37	0.08	0.17	0.65	0.10	0.20
HRC9	0.68	0.06	0.11	0.17	0.07	0.14	0.48	0.06	0.11	0.63	0.06	0.12	0.13	0.05	0.10	0.35	0.09	0.17
HRC10	0.75	0.05	0.10	0.41	0.07	0.14	0.63	0.05	0.09	0.73	0.05	0.10	0.41	0.08	0.15	0.62	0.08	0.16

Table 6.38 – Mean (μ), Confidence Interval at 95% (CI) and Standard Deviation (σ) of all SRCs per HRC (IPPP GOP1)

If one considers the grading system based on the five-point scale, then, for IBBP GOP1, two sequences maintain the "quality category" regardless of the reference bit rate: SRC1 (Akiyo) with "Excellent," and SRC6 (sil) with "Good." Analyzing SRC2 (fot), regarding 512kbps, HRC4 ("Mux. Bit - 512kbps") the scores are "Poor" and the remaining HRCs scores are "Fair" (HRC2, HRC6, HRC8, and HRC10). As for 256kbps results (SRC2), the scores of HRC7 ("Mux SSIM - 256kbps") are "Fair" and the rest of the HRCs scores (HRC1, HRC3, HRC5, and HRC9) are "Poor". Regarding SRC3, all results are "Good", for 256kbps and 512kbps, with the exception of HRC2 (H.264 JM - 512kbps) that is "Excellent". For the last two SRCs, SRC4 and SRC5, at 512kbps, the score is equal for all HRCs, and respectively "Excellent" and "Good". As for SRC4 and SRC5 scores at 256kbps, almost all HRCs show a "Good" score. The exception is HRC1 whose score is "Excellent" for SRC4, and "Fair" for SRC5. Using the five-point scale, it can be seen the effect of joint coding to obtain a higher degree of uniformization of the image quality of the video sequences.

Finally, let us analyse the results of the various SRCs for the IBBP GOP1 test session. The interval of variation of the scores, for the distinct PVSs, grouped by SRC, differs quite a lot. This range is smaller when encoding video sequences with a lower spatio-temporal complexity. As an example, this result was observed in the Akiyo and Mad sequences. As for the video sequences with higher spatio-temporal complexity characteristics, the range of deviation is much higher (for example, Mcl or Fot video sequences). The degree of differences is bigger between HRC associated with 256kbps than with HRC that are associated with 512kbps. As for the sequences, the Akiyo sequence presents the scores nearest its mean value, and the Fot sequence presents scores more distant from its mean value.

For the test session IPPP GOP1, five video sources have better subjective scores compared with the independent coding: SRC2 (HRC5, HRC6, HRC7, HRC8), SRC3 (HRC5), SRC4 (HRC5, HRC7), SRC5 (HRC3 to HRC10) and SRC6 (HRC6, HRC7, HRC8, HRC10). Comparing the scores of HRCs, in the two test sessions, in both cases, best results were obtained for Mux PSNR and for Mux SSIM. The number of SRC that has improved is slightly higher in the test session IPPP GOP1: HRC5 (Mux PSNR - 256kbps) and HRC7 (Mux SSIM - 256kbps).

Using the five-point scale, for IPPP GOP1, no sequence displays the same "quality category" for all the HRCs. The results of the different video sources can be analyzed by grouping the HRC according to the value of the reference bit rate. For 256kbps, it can be seen that two SRCs kept the same evaluation for the five HRCs (HRC1, HRC3, HRC5, HRC7, and HRC9): SRC1 ("Good"), and SRC3 ("Fair"). For 256kbps, SRC4 and SRC5, most of the HRCs have the same category: "Good" and "Bad" respectively. The exception is HRC3 (Mux Bit) whose results are balanced: SRC4 ("Fair") and SRC5 ("Poor"). Finally, in the 256kbps analysis, SRC2 and SRC6

are generally classified as "Bad" and "Poor.” Exceptions to this classification are HRC5 (SRC2 “Poor") and HRC7 (SRC2 "Poor" and SRC6 "Fair").

Regarding the results of the test session "IPPP GOP1," for 512 kbps, they differ greatly from the results observed so far. The scores of independent coding (HRC2) are quite diverse: SRC1 and SRC4 are "Excellent," SRC3 is classified as "Good," SRC2 and SRC6 are classified as "Fair" and SRC5 as "Poor." Comparing HRC4 (Mux Bit) with HRC2, it is found that it presents the highest number of changes in the classification: four SRC have their classification decreased by one level (SRC1, SRC2, SRC3, and SRC4), and SRC5 increases its classification one level, from "Poor" to "Fair." These results confirm the observations made during the analysis of objective results regarding Mux Bit. The best results are presented in HRC8 (Mux SSIM) and HRC10 (Mux PSPNR). As for the results of HRC8, SRC6 increases its classification from "Fair" to "Good," four SRCs maintain the classification value (SRC1, SRC2, SRC3, SRC5) and SRC4 lowers its rating from "Excellent" to "Good." As for the results of HRC10, two SRCs increase their classification one level (SRC5 from "Poor" to "Fair," and SRC6 from "Fair" to "Good"), two SRCs maintain their classification (SRC2, and SRC3), and two SRCs decrease their classification one level (SRC1 and SRC4, from "Excellent" to "Good"). Finally, regarding the results of HRC6, four SRCs maintain their classification (SRC2, SRC3, SRC5, and SRC6), and two SRCs decrease their classification one level (SRC1 and SRC4, from "Excellent" to "Good").

Finally, the dispersion of the scores is greater during the test session IPPP GOP1 compared with the test session IBBP GOP1. In each of the test sessions, the dispersion of the scores varied according to the characteristics of video sequence and the value of the bit rate reference. The scores tend to be more deviated from the mean, in PVSs with the reference bit rate of 256kbps. Analyzing the dispersion in each sequence, there is a higher concentration of scores in the sequences Akiyo, Hall or MAD, and a higher dispersion in the sequences Mcl and Fot. This information shows that observers have more agreement on the quality assessment of video sequences with lower spatio-temporal complexity than on videos with higher spatio-temporal complexity.

6.5 Two-pass Video Coding incorporating Perceptual Metrics

For non-real time applications, such as digital storage applications, it can be desirable to achieve better visual quality by allowing larger bit rate variation in adjacent frames, at the cost of higher computational complexity. One solution is to implement a multi-pass strategy. The fundamental principle of this approach is in the first pass to encode the whole or part of the sequence using a fixed quantisation parameter or to encode at CBR. In this first passage, the encoder generates

data about the encoding statistics such as the coding complexity of the frames. Next, a coding model incorporating the collected coding statistcs of the first passage is built. This model is used to determine the quantisation parameters in the second passage in order to improve the picture quality of the video bitstream. In this section, a proposal for a two-pass algorithm will be presented that could be further developed in future research.

Teixeira et al. introduced a two-step MPEG-2 video system taking variance as the scene complexity measure ([462],[463]). Westerink et al. ([461]) introduced a two-pass rate control algorithm using the quantisation scale, spatial activity and temporal activity of each frame to build an R-Q model. Yu et al. ([466]), proposed to compute for each frame the R-Q function for all possible quantisation parameters. Analyzing results from first-pass, three optimal quantisation factors for the three different picture types are determined based on the MPEG2 TM5 rate control model. Then, for each frame, the quantisation parameter is tuned based on the R-Q function. As all quantisation parameters for encoding the second pass are previously determined, the encoder cannot make an adjustment according to the actual consumption of bits in the second encoding pass. Thus, a large gap between the bit rate and bit rate target may occur. The Lie et al. ([512]) method is based on analyzing window segments of the video sequence. The analysis of the first-pass statistics is performed so that models of Rate versus Lagrange's multiplier and Distortion versus Lagrange multiplier are built. These models are used to adjust the quantisation parameter for each macroblock. Given that the two-pass encoding is based on limited window segment, there is partial knowledge of the future video frame's characteristics. Thus, it is impossible to accurately distribute the bit rate according to the characteristic of the complete video sequence. Kwon et al. ([513]) propose a GOP-level based two-pass rate control algorithm. Results show a higher coding efficiency and smoother video quality compared to JM H.264/AVC CBR rate control. Again, constant quality video coding is not achieved for the complete video sequence. Que et al. [514] propose a method to extract statistics in the first-pass, such as intra-mb distribution information, to detect scene cuts and allow performing GOPs regrouping. This method employs "global complexity measure" and PSNR information in the bit rate allocation. Although it is targeting global optimization, the complexity metric is still based on the old TM5 definition and the traditional PSNR. Furthermore, he proposes a method without a mechanism to examine, in the second-pass encoding, the bit control status and uses the preset quantisation parameter for each frame. Huang et al. propose in ([515]) an approach to extract and perform a statistical analysis of the integer transform coefficients, obtained in the first-pass. With this information, a R-D model is built so that quantisation parameters can be optimized for all the frames. The aim is to enhance the consistency of video quality during the entire video sequence. Still, this approach experiences high bit rate variance because it employs

an R-D model obtained in the first passage to compute the quantisation parameters in the second passage without taking care in the second pass of the bit rate dynamics. Consequently, each frame will spread the bit rate divergence as no control mechanism exists to verify the actual status of the second-pass encoding buffer.

All the above proposals are based on traditional image quality metrics. The incorporation of perceptual metrics in a two-pass or multi-pass strategy improves the perceptually uniform distortion within an image, allowing the smoothing of temporal fluctuations in image quality and to enhance the image quality among video programmes. In this section, a proposal will be presented that benefits from work describe in this thesis and that could improve overall perceptual image quality results. The algorithm allows the quantisation parameter to be varied locally based on the subjective quality criterion. The cost of image coding is distributed according to a subjective criterion making it possible to allocate more bits to the areas most sensitive to coding errors from the perspective of the viewer.

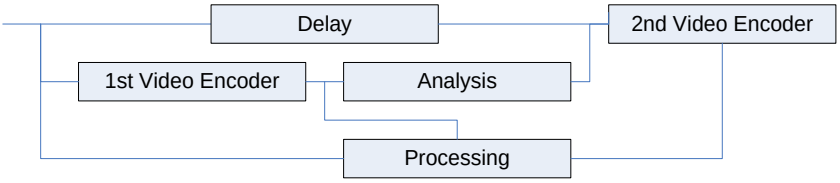


Figure 6.18 – Block Diagram for Two-pass Video Coding

The original image is sent simultaneously to a Delay Block, the entry of a first video encoder, and the entry of a processing block (Figure 6.18). The first video encoder sends a local decoded image to the processing block, and data relating to the performed compression to an analysis block. The output of the analysis block and the processing block are connected with the second video encoder. The first video encoder compresses the source image using a fixed quantiser. The output of the first video encoder provides a local decoded image. The processing block compares the source image with the degraded image results from the first-pass. The processing block provides the second video encoder with a visual distortion sensitivity gray image, where the intensity of the each pixel is proportional to the visibility of the error in the original image. The second video encoder also receives the information generated from the analysis block. The delay block allows controlling when an original image is sent to the second video encoder. The delay time is a function of the number of images that need to be analysed before the second video encoder starts working. The second video encoder, therefore, encodes these delayed images as a function of the information generated from the analysis block and the processing block. Thus, the second video encoder can efficiently allocate bits to address different demands such as to allocate bits to encode the motion information or to encode the texture information.

Rate control needs to decide how to allocate available bits such that one parameter may benefit at the cost of another parameter. For example, to allocate more bits to provide accurate motion information versus allocating more bits to provide texture information. With information about which areas in the current frame are particularly important, or worth more bits, the encoder can allocate of the available bits more efficiently.

The goal of the rate control algorithm is to modulate the quantisation parameter according to a subjective quality criterion. For a given bit rate, the overall quality of the image should be improved, by degrading areas where viewers are less sensitive to the quality of the image and by improving the areas where the viewers are more sensitive to image quality. A psycho-visual model gives subjective quality criterion. Typically, this type of model carries out a subjective evaluation on a degraded image in relation to a reference image, typically the source image, and determines the position of the most sensitive areas to errors. The degraded image is usually a local decoded image obtained during the first encoding pass. After obtaining the two images, the degraded and the reference image, one possible psycho-visual model is a JND type, mapping the error into a subjective perception map with values representing the level of perception of the error by the human eye. The subjective perception map can be determined on a macroblock basis as follows:

$$JND_{i,MB}(k,j) = \frac{\sum_{l=0}^{Nlines-1}\sum_{c=0}^{Ncol-1} JND_{pixel}(k*Nlines+l,\ j*Ncol+c)}{Nlines*Ncol} \tag{6.62}$$

where Nlines is the number of lines of the macroblock, Ncol is the number of pixels per line, k and j respectively represent the column number and the line number of the macroblock MB(k,j) in the image i. A similar JND Map has been presented in Figure 2.19. A mean value of the JND over the image i, $\overline{JND_i}$, is determined to measure the overall image quality after quantisation:

$$\overline{JND_i} = \frac{\sum_{u=0}^{SI-1}\sum_{v=0}^{MB-1} JND_{i,MB}(u,v)}{SI*MB} \tag{6.63}$$

where MB is the number of macroblocks per slice, and SI is the number of slices in an image. To determine the scene changes of the complete sequence it is enough just to compute the JND difference map as follows:

$$\Delta_{JND_i} = \frac{\sum_{u=0}^{SI-1}\sum_{v=0}^{MB-1} \left| JND_{i,MB}(u,v) - JND_{i-1,MB}(u,v) \right|}{SI*MB} \tag{6.64}$$

To detect scene cuts we only need to compare Δ_{JND_i} with the mean value of the differences of all coded images $\Delta_{\overline{JND}}$ as follows:

$$\Delta_{JND_i} > \beta \times \Delta_{\overline{JND}} \tag{6.65}$$

where β is a constant (typical value should be above 3.0). As scene changes are relevant for images of type P and B, evaluating this type of image is enough. Next, the average value of the $\overline{JND}$ over a group of images is determined. The number of images, D, is an input parameter and associated with the delay block.

$$\overline{JND_D} = \frac{\sum_{u=0}^{D-1} \overline{JND_u}}{D} \tag{6.66}$$

If the purpose is to increase the uniformity of image quality during the entire video sequence, then D should be equal to the number of frames in the video sequence. In a near real time scenario, D can be set to the number of image in a GOP, to avoid frequent changes in the quantisation parameter. After having calculated the value of $\overline{JND_D}$, a comparison is made by determining the ratio between the JND of each macroblock in the image with the mean value of JND for the D images.

$$\alpha = \frac{JND_{MB}(k, j)}{\overline{JND_D}} \tag{6.67}$$

If the value JND_{MB} of the macroblock (k,j) is near to the value $\overline{JND}$, then it is not necessary to perform an adjustment. If the value JND_{MB} is bigger than the average value of JND then viewers will be more sensitive to coding error in this area, and the value of the quantification parameter should be lowered. Otherwise, if JND_{MB} is lower than $\overline{JND_D}$, then the viewer will be less sensitive to coding errors in this area and thus we may increase the quantisation parameter.

The last step is to bind the coefficient α between the lower and upper bound limits, $\alpha_{\min}$ and $\alpha_{\max}$, to avoid big changes of the quantisation parameter. In fact, great changes can cause instability in the rate control algorithm. Thus, the quantisation parameter should be adjusted as follows:

$$Q_{MB} = \frac{Q}{\max(\alpha_{\min}, \min(\alpha, \alpha_{\max}))} \tag{6.68}$$

where Q_{MB} is the adjusted quantisation parameter to the macroblock MB taking in consideration the perceptual weighting criterion and Q is the initial quantisation parameter that was calculated by the rate control algorithm.

When the goal is to increase the uniformity of the image quality during the entire video sequence then, for each GOP, an adjustment in the GOP-level bit allocation step is proposed regarding Equation (4.60). After encoding the j[th] picture in the i[th] GOP, the total bits for the remaining pictures in the i[th] GOP can be computed as follows

$$B_i(j)=\begin{cases}\dfrac{\overline{JND_i^{GOP}}}{\overline{JND_D}}\dfrac{R_i(j)}{f}\times N_i - V_i(j) & j=1\\ B_i(j-1)+\dfrac{\overline{JND_i^{GOP}}}{\overline{JND_D}}\dfrac{R_i(j)-R_i(j-1)}{f}\times(N_i-j+1)-b_i(j-1) & j=2,3,...,N_i\end{cases} \tag{6.69}$$

where f is the predefined coding frame rate. N_i is the total number of pictures in the i[th] GOP. $R_i(j)$ and $V_i(j)$ are the instant available bit rate and the occupancy of the virtual buffer, respectively, when the j[th] picture in the i[th] GOP is coded, and $b_i(j\text{-}1)$ is the actual generated bits in the (j-1)[th] picture ([339]). $\overline{JND_i^{GOP}}$ is the average value of JND for the i[th] GOP.

This is just a proposal for future research. This proposal considerably avoids alterations of the quantisation parameter, and at the same time it promotes the increase of subjective quality in the sensitive areas. The integration of this proposal in the different joint coding systems that were presented in this Chapter could additionally enhance the video quality within each video sequence.

6.6 Summary

In this Chapter, the concepts of statistical multiplexing and joint video encoding were introduced. In joint video encoding systems, a common bit budget is divided between video programmes according to a criterion by a joint rate controller. The principal difficulty of these systems is how to share the bandwidth between the video programmes that use the same channel bandwidth. Numerous solutions have been presented in the literature and are described in section 6.2. A popular approach is to use a look-ahead scheme to collect statistics about the coding complexity of the video programmes. After, when the coding process starts, bandwidth is allocated based on the statistics obtained in the first step. These systems use a look-ahead window, and thus the system delay is increased. Another current alternative, based on a RD model, is to model the video encoder performance and the coding complexity. Then, based on the RD models bandwidth is shared. A linear model or an exponential RD model have been

used. Typically, MSE or PSNR is used to assess distortion / picture quality. As no standard methodology exists, simulation's settings differ quite a lot (video spatial and temporal resolution, GOP structure, motion parameters, etc). Thus, it is difficult to compare results.

Based on the bit rate modelling of previous Chapter, rate-quantisation functions are exploited for allocating bandwidth between the different programmes and to estimate quantisation parameters. A look-ahead process is used with a window of size one GOP. Distortion is measured using perceptual metrics and frames coded with I-slices, P-slices and B-slices pictures are modelled using perceptual metrics. In general, in the literature, simulation results are presented with only Intra and Predicted slices. In the present work, GOP structures using B-slices were also used. Based on results regarding the best ways to model D-QP, R-QP and R-D using perceptual metrics, a joint encoding scheme is proposed. The goal is to obtain uniform picture quality. It should be noted that the process is more complex than a normal CBR channel with CBR video programmes being sent or a joint video encoding system using SAD as SSIM requires more calculations. A full integration of SSIM in the video encoder could improve the video encoder coding performance. In recent years, several proposals have been made to replace SAD by SSIM in RDO process in the motion estimation and mode decision process (Chapter 5). A full SSIM video encoder implies a higher complexity as the number of operations to compute SSIM are higher than to compute SAD (Table 6.39). For example, SSIM needs more than 5 times more operations than SAD to perform motion estimation on a 8×8 block.

Operation	SAD	SSIM
Add	$2 \times N^2 - 1$	$7 \times N^2 + 3$
Multiply	0	$3 \times N^2 + 17$
Total operations per block	$2 \times N^2 - 1$	$10 \times N^2 + 20$

Table 6.39 – Numbers of Operations of SAD and SSIM for Motion Block Size (N×N)

The performance of the proposed algorithm was evaluated through simulations over 2, 3 and 6 video programmes. From video quality perspective, the proposed scheme has been compared with independent video rate control (more constant quality than independent rate control over the programmes and in some cases a higher average picture quality value). Figure 6.19 and Figure 6.20 illustrate the gain in perceptual image quality when three sources are jointly encoded, and SSIM is used in the joint encoding process. Improvement in image quality is clear for the Football sequence while Akiyo suffers a small decrease in picture quality.

Figure 6.19 – Akiyo (frame 35), AAC, 256 kbps (from left to right - independent coding, joint coding SSIM)

Figure 6.20 – Football (frame 35), AAC, 256 kbps (from left to right - independent coding, joint coding SSIM)

Generally, a uniform distribution of distortion in the different video sequences was obtained. Overall results of image quality differed when measured by PSNR or SSIM. Thus, tests were carried out, using the SAMVIQ method to assess subjective image quality. Again, results confirm that joint coding methodology based on perceptual metrics can obtain a higher degree of uniform image quality among encoded video sequences, and that the mean perceived quality improves. Finally, a proposal is made, in terms of future research, for a two-pass rate control algorithm. In the first-pass, a JND psycho-visual model is used to represent the error as a subjective perception map and thus assess scene complexity. In the second-pass the quantisation parameter is modulated according to the subjective quality criterion. The integration of this proposal is on-going.

Chapter 7. Conclusions

The main goal of this dissertation was the development of novel techniques that allow to incorporate perceptual picture quality metrics into joint video encoding systems. In order to achieve this goal the work described in this thesis focused on the development of techniques in two areas: RD modeling using perceptual picture quality metrics (Chapter 5) and joint video encoding techniques (Chapter 6).

In Chapter 2, an introduction to digital video quality was provided. There are two distinct classes of methods available to perform video quality assessment ([67],[68]): subjective and objective measurement's methods. A summary of the different methods to classify objective quality metrics was presented giving particular emphasis to their main characteristics. Objective video quality metrics can be categorised in three classes: Full Reference (FR), Reduced Reference (RR) and No Reference (NR) ([108],[130],[131],[211],[212]). In this work, the objective metrics used are Full Reference only. It was found that FR methods outperform RR and NR methods. However, it is hard to have full access to the source programme all time. The field of NR image quality metrics still remains largely unexplored to date and has received increasing attention ([516],[517]). A very recent NR metric, based on the concept of structural activity (SA) together with a model of SA indicator in a new framework for NR image quality assessment have been recently proposed ([516]). According to the authors, there are still a number of topics that need further research such as how to incorporate the appropriate HVS properties at acceptable complexity or the development of other effective implementations of SA indicator or hybrid quality measures to extend the scope to more distortion types and multiple distortions. Although NR models could solve this problem, more research of NR methods is still required to reach the same level of prediction accuracy as the FR and RR methods ([212]). Traditional FR metrics, such as MSE and PSNR, are a simple and fast way to predict video quality. Besides PSNR, two FR approaches for quality assessment, based on JND and on structural similarity (SSIM), were presented and discussed regarding their concept, implementation and meaning. In contrast with traditional metrics, they used mechanisms to incorporate HVS or the perceptual effects of video degradation. As a result, they allow a more refined prediction of the level of degradation that a signal can suffer until a human observer notice it ([212]).

The Chapter 3 provided a brief introduction to some of the most well-known international video coding standards (MPEG-1 [221], MPEG-2 [2], MPEG-4 [227], and H.264/AVC [6]). Special attention is devoted to H.264/AVC that adopts many new video coding tools such as intra prediction, integer transform, enhance inter prediction, context-based entropy coding, and deblocking filter. These new technical developments mean that H.264/AVC achieves a key breakthrough on Rate-Distortion performance. In this work, H.264/AVC is the platform where the proposed techniques will be integrated. Video coding standards have their own recommendation on rate control as an informative part based on the work developed during the development phase. The H.264/AVC JM describes a rate control algorithm that employs a rate-distortion (RD) optimisation technique and is compliant to the H.264/AVC standard HRD.

Recent video standards such as MPEGx or H.26x standards aim to facilitate interoperability and data exchange among different products or services ([2],[6],[214],[215],[221]). In order to achieve these goals, they specify the requirements imposed on the complete bitstream syntax and decoders. The standardisation of the decoders enabled independent implementations, from different software and hardware manufacturers, and ensured that those implementations will be interoperable. Video standards do not normally define how to perform rate control. Nevertheless, during its development process, algorithms were verified through tests, simulations, and verification models. To allow testing and to perform simulations using a common set of encoder routines, both MPEGx and H.26x set up a sequence of test models as an informative tool (non-normative tool). Each test model normally suggests a rate control method during its development phase, e.g. TM5 for MPEG-2 ([19]), TMN8 for H.263 ([20]), and VM8 for MPEG-4 ([308]), etc. An improved rate control method based on VM8, supporting rate distortion optimisation (RDO), has been adopted by H.264/AVC JM test model ([169],[170]). The Chapter 4 focused in this rate control algorithms.These rate control algorithms do not aim to deliver an optimal solution. The adopted rate control algorithms are competitive in R-D performance, with acceptable computational complexities, and are flexible in terms of adaptation capacities regarding different video sources. Thus, it is valuable to review rate control algorithms as they incorporate the progresses obtained in recently developed rate control techniques. Moreover, the level of performance obtainable by these methods serves as a point of comparison for future research and the development of rate control methods. In general, a rate control algorithm has two steps: resources allocation and computing of the quantisation parameter. The first step can be performed among different video objects (the rate control of multiple video objects), different frames (the rate control of single sequence) or different sequences (the rate control of joint sequences, which is the focus of the present work). In the

second step, the coding mode is selected and the quantisation parameter computed, usually based on a R-D (Rate-Distortion) model and RDO (Rate Distortion Optimization) process.

In Chapter 5, previous research conducted in the field of Rate Control optimization was examined, particularly R-D modeling and the developing of R-D functions for the rate control of joint video sequences using perceptual quality metrics. Extensive experiments on a large number of video sequences were performed, their statistics studied and a Rate-Quantization (R-Q) model and Distortion-Quantization (D-Q) model derived for modeling the R-D relation in H.264/AVC. Rate-distortion (R-D) based methods are often implemented to enhance and stabilize video quality. The MSE and PSNR are the most used video quality metrics, given their little computational complexity, regardless of their limitations. In our experiences, the image quality metrics was extended to include SSIM and PSPNR as MSE and PSNR do not correlate well with perceived quality. Experimental results show that quadratic function is a good solution to model R-D using SSIM or PSPNR. Simulations were performed for different GOP Patterns, video test sequences, and coding setup.

In Chapter 6, novel joint video encoding schemes based on perceptual RD models are introduced. The key idea behind joint video encoding is to allocate bits to every programme according to their time-varying content complexity. The impact on picture quality has been measured by three different criteria: PSNR, PSPNR and SSIM. This differs from traditional analysis where the visually impact of the encoded test video sequences is frequently measured only by PSNR or MSE metrics. Four joint coding algorithms have been implemented and assessed. In two of the propose algorithms, it was integrated perceptual metrics (SSIM and PSPNR). Results show that by transferring bits between streams with less coding complexity into video programmes with high coding complexity that the maximum level of distortion can be decreased, more homogeneous video quality sequences can be obtained. Reallocation of bits is obtained without compromising the overall quality of video programmes. This proposal can be easily being extended to support encoded video stream with different frame rates, spatial resolutions, or other's formats.

As video programmes are in the end to be view by human beings, the most consistent way of assessing the quality of video is a subjective evaluation. This assessment technique is can be rather complex, time consuming, uncertain (when a subjective test session is design for the first time, frequently results cannot be exploited due to errors in the test design) and costly (because of the human resources involved) task. However, we felt that it was rather important to correlate the obtained results with a subjective metrics and human evaluation system. ITU-R Rec. 500 [70] identifies the need to translate the quality adjectives into the language of the country where each subjective test session is performed (Excellent, Good, Fair, Poor, Bad). Nevertheless,

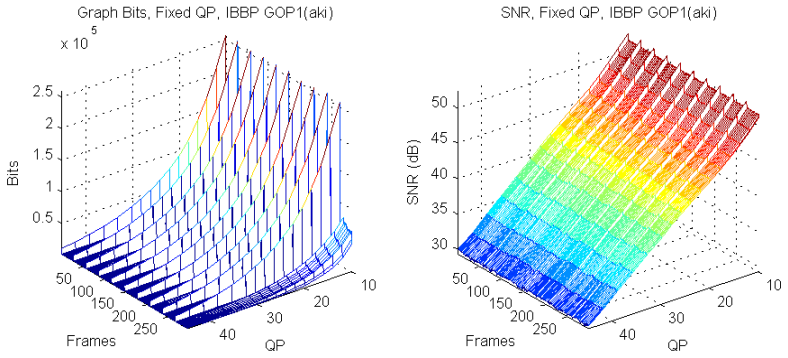
according to ITU-R, it is also accepted that the translation to different languages presents a small bias due to the diverse connotation that each language gives to the translated terms. Results from subjective test sessions validate objective results of Mux SSIM and Mux PSNR results. In summary, the main contributions of this thesis are as follows: Modelling of R-D of H.264/AVC using perceptual metrics to assess the distortion; development of joint source coding algorithms for controlling a statistical multiplexing process of different streams into a fixed bandwidth channel that incorporate perceptual information; and study how joint coding results correlate with subjective quality assessment.

Current work involves the integration of the presented full two-pass algorithm with a H.264/AVC SSIM-RDO encoder. To limited the expected complexity, fast SSIM implementations and early termination techniques are being evaluated to reduce complexity and simulations time.

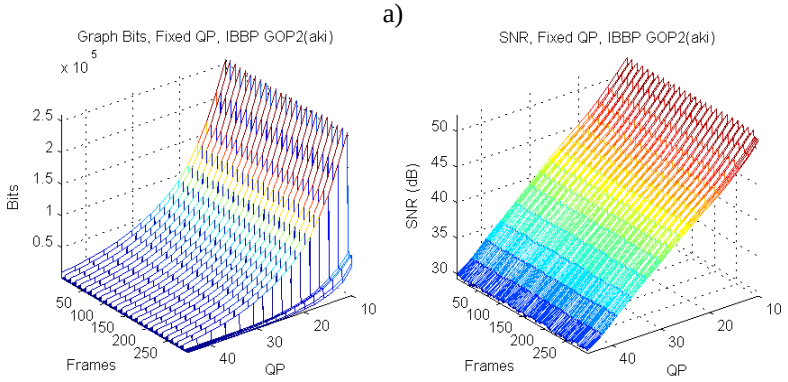
Annex

Annex A. Picture Quality Metrics as a function of Quantisation

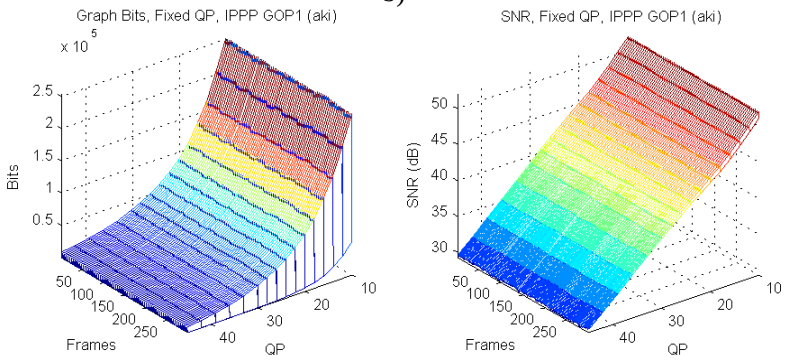
A.1 Frame Size and Picture Quality (SNR) versus QP



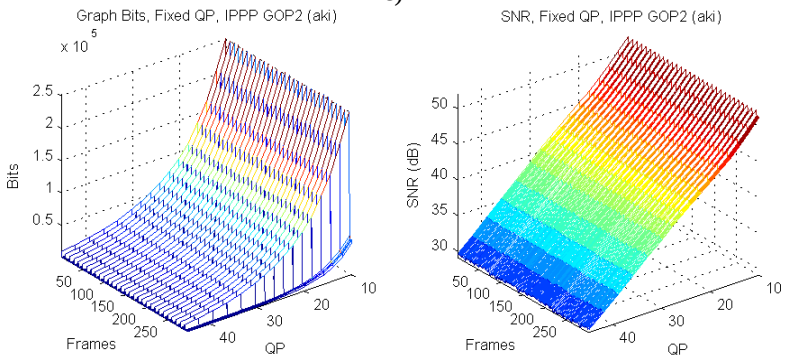
a)



b)



c)



d)

Figure A.1 – Bits and SNR for H.264 Akiyo video stream encoded with fixed QP and with different GOP Patterns: IBBP GOP1 (a), IBBP GOP2 (b), IPPP GOP1 (c), and IPPP GOP2 (d)

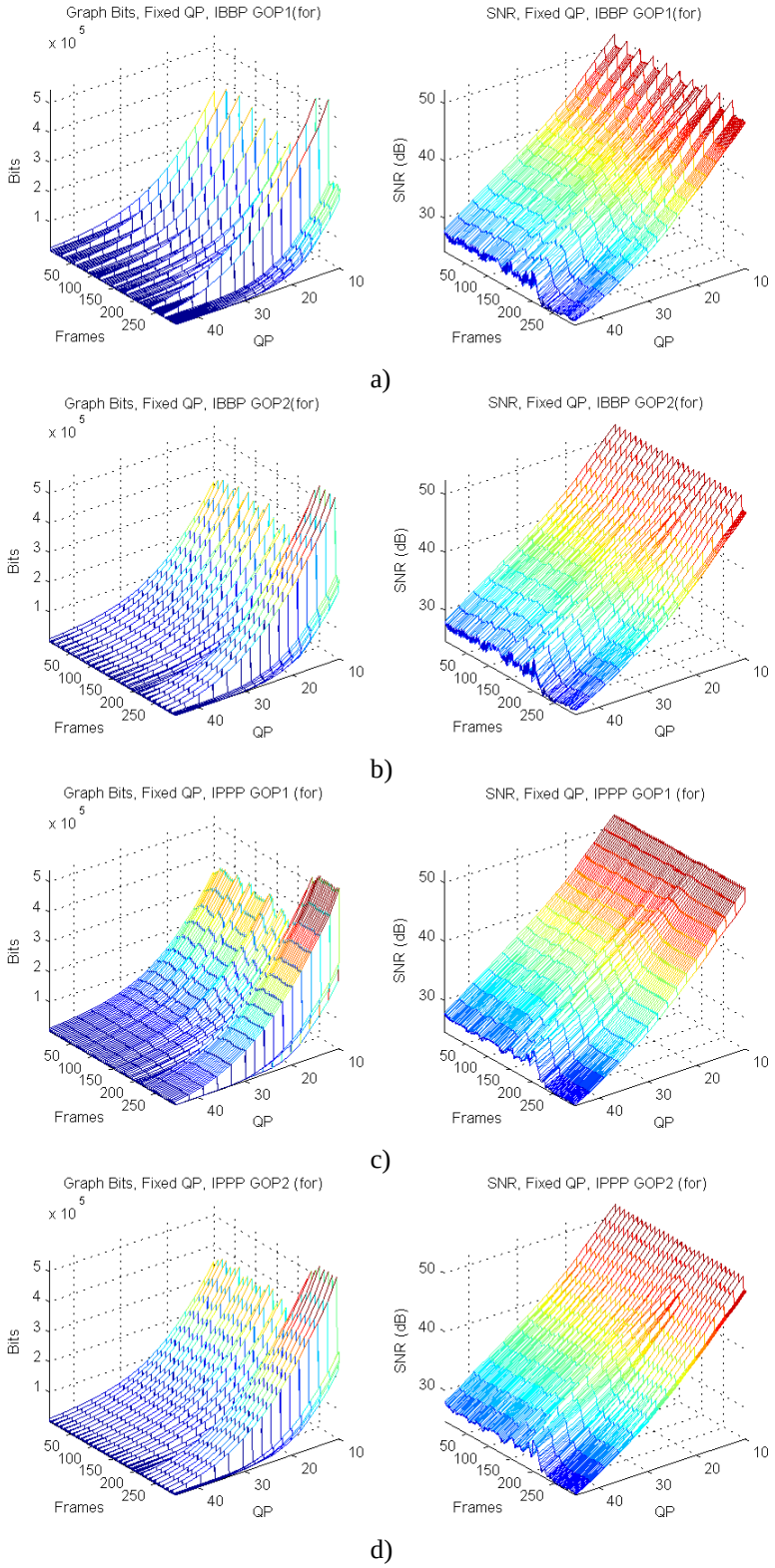


Figure A.2 – Bits and SNR for H.264 Foreman video stream encoded with fixed QP and with different GOP Patterns: IBBP GOP1 (a), IBBP GOP2 (b), IPPP GOP1 (c), and IPPP GOP2 (d)

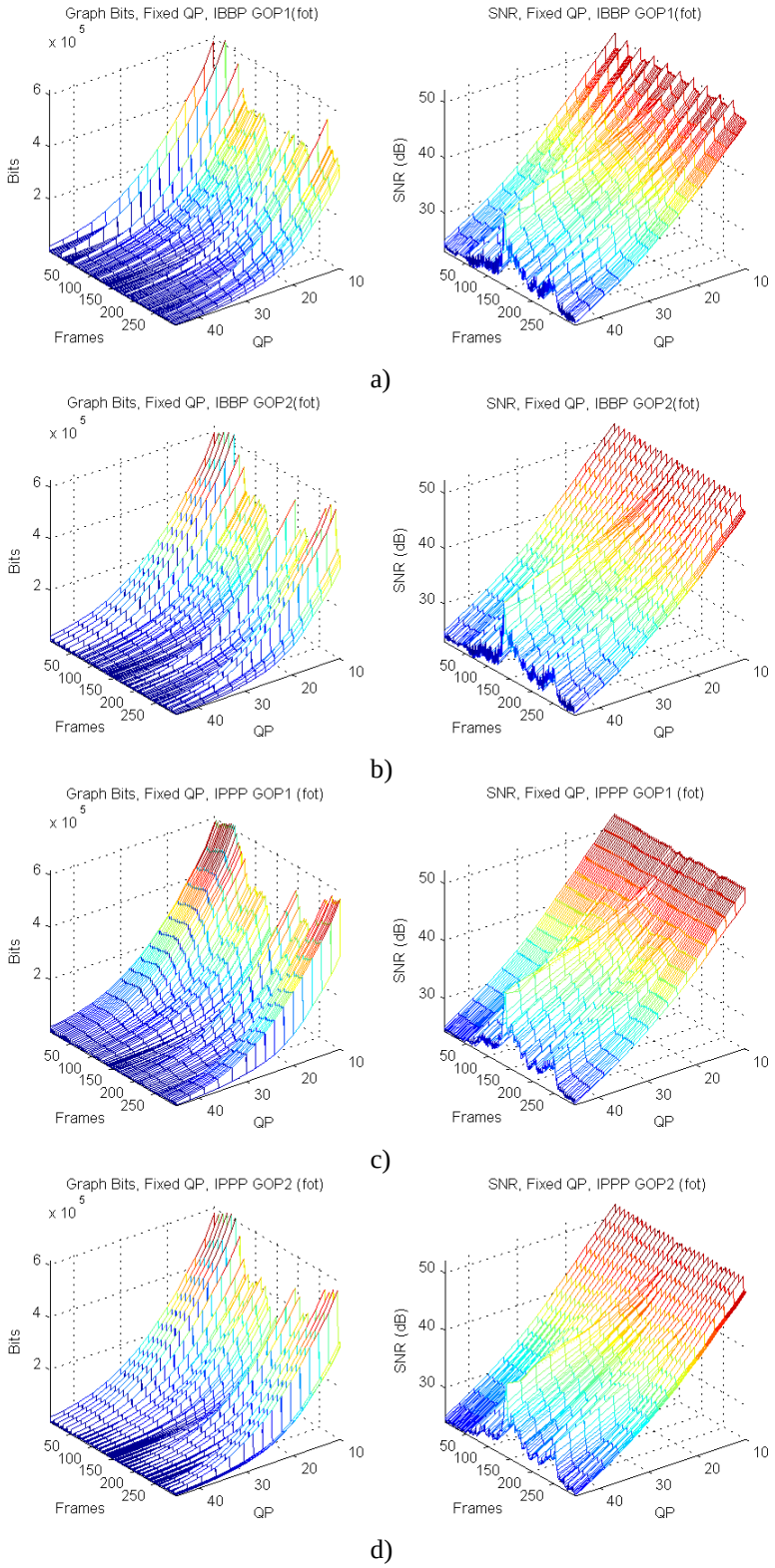


Figure A.3 – Bits and SNR for H.264 Football video stream encoded with fixed QP and with different GOP Patterns: IBBP GOP1 (a), IBBP GOP2 (b), IPPP GOP1 (c), and IPPP GOP2 (d)

A.2 Picture Quality Metrics (PSNR, PSPNR, SAD JND and SSIM) as a function of Quantisation

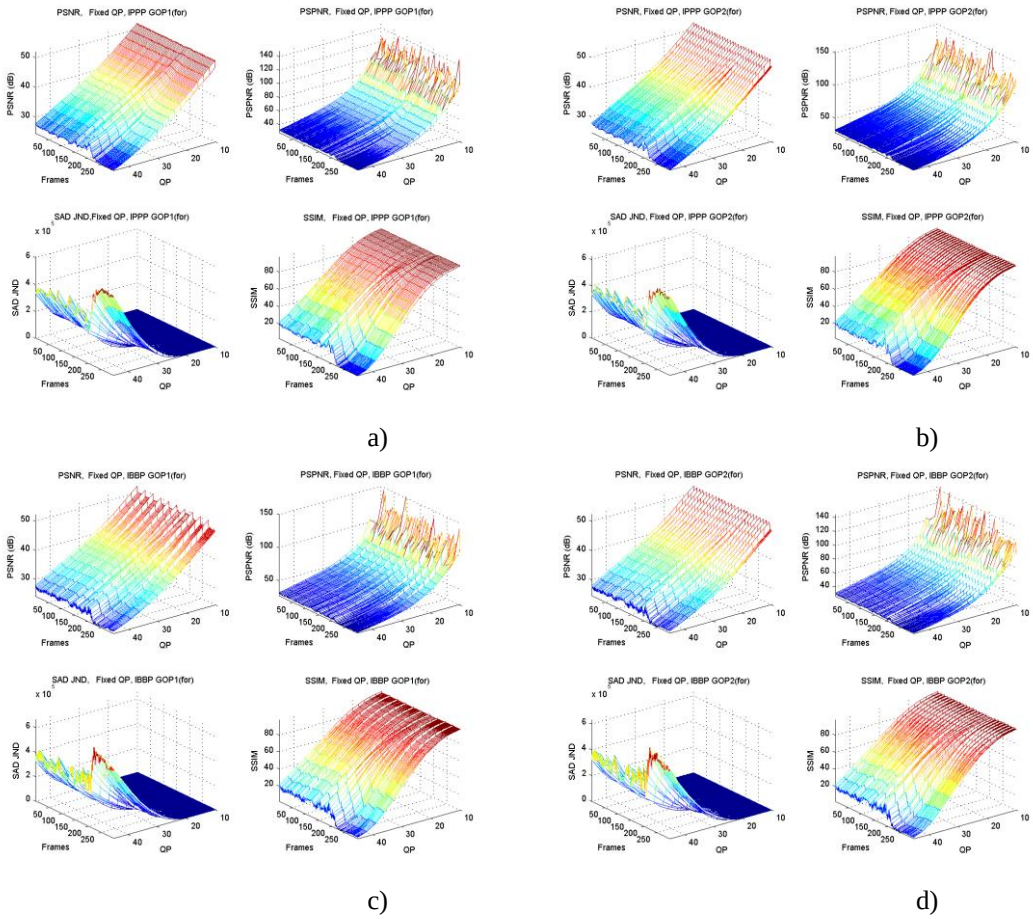


Figure A.4 – Picture quality metrics (PSNR, PSPNR, SAD JND and SSIM) as a function of quantisation for video H.264 video sequence Foreman with different GOP structures: IPPP GOP1 (a), IPPP GOP2 (b), IBBP GOP1 (c), and IBBP GOP2 (d)

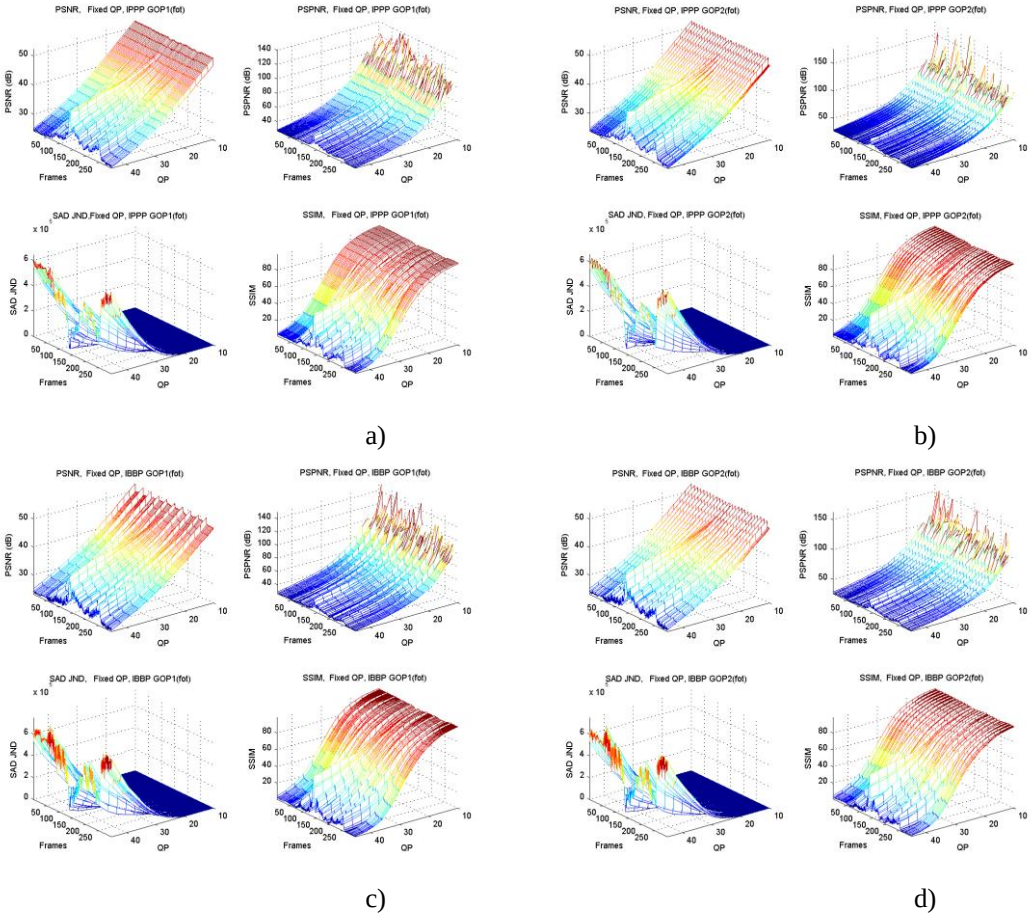


Figure A.5 – Picture quality metrics (PSNR, PSPNR, SAD JND and SSIM) as a function of quantisation for video H.264 video sequence Football with different GOP structures: IPPP GOP1 (a), IPPP GOP2 (b), IBBP GOP1 (c), and IBBP GOP2 (d)

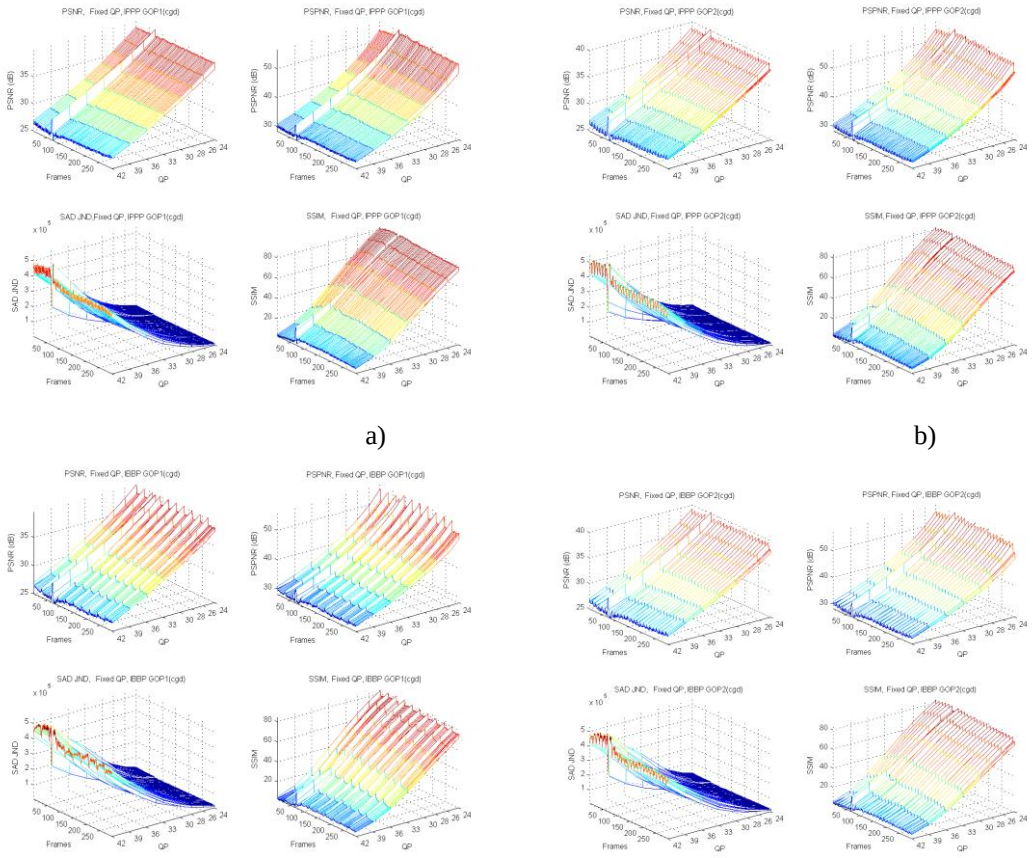


Figure A.6 – Picture quality metrics (PSNR, PSPNR, SAD JND and SSIM) as a function of quantisation for video H.264 video sequence CoastGuard with different GOP structures: IPPP GOP1 (a), IPPP GOP2 (b), IBBP GOP1 (c), and IBBP GOP2 (d)

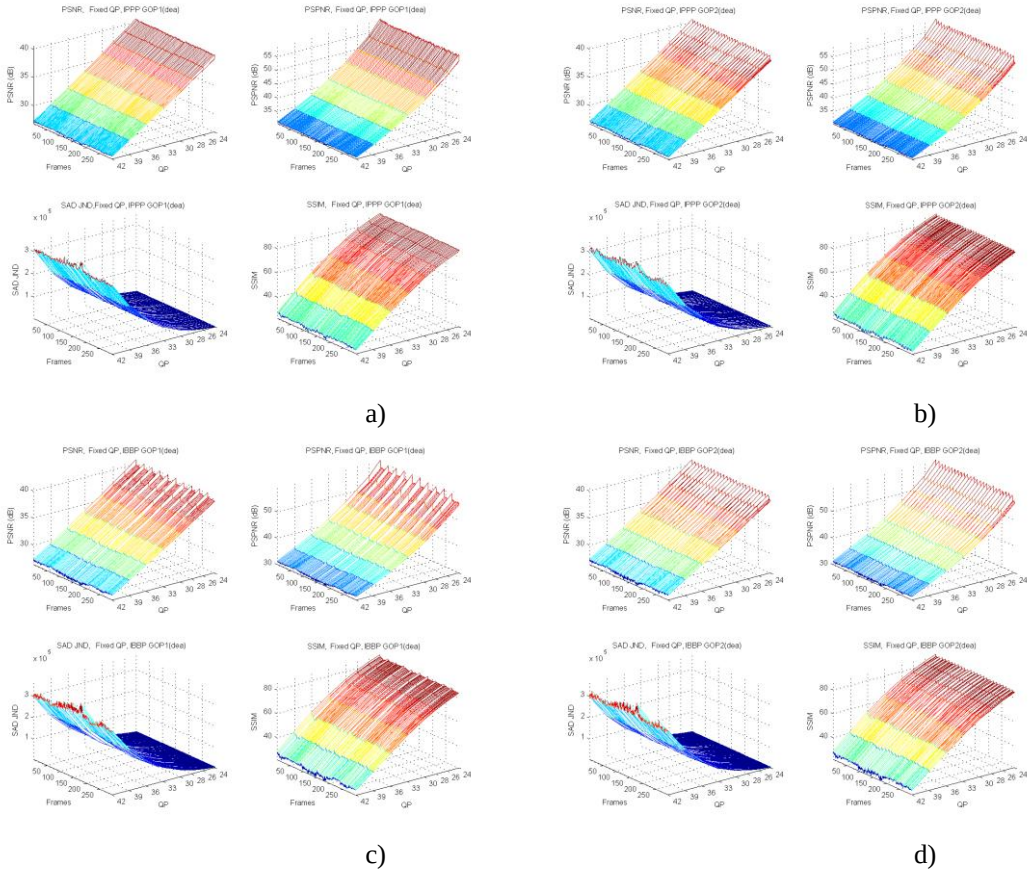


Figure A.7 – Picture quality metrics (PSNR, PSPNR, SAD JND and SSIM) as a function of quantisation for video H.264 video sequence Deadline with different GOP structures: IPPP GOP1 (a), IPPP GOP2 (b), IBBP GOP1 (c), and IBBP GOP2 (d)

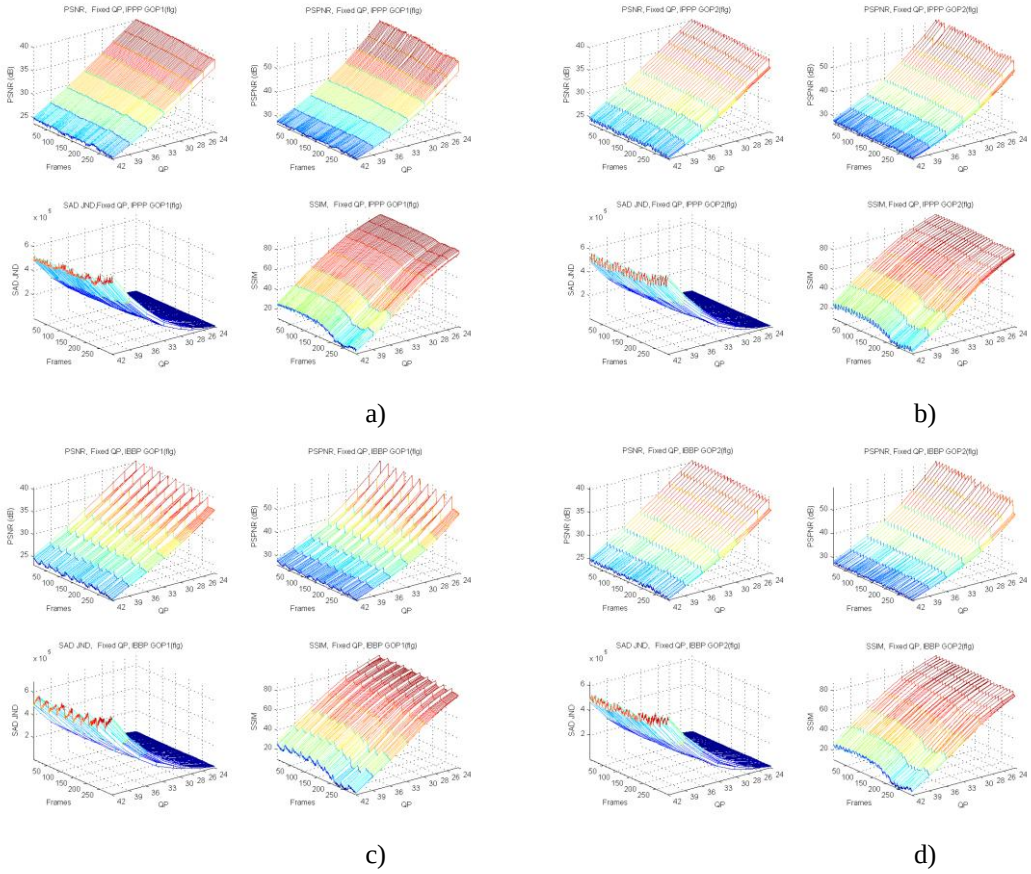


Figure A.8 – Picture quality metrics (PSNR, PSPNR, SAD JND and SSIM) as a function of quantisation for video H.264 video sequence Flower Garden with different GOP structures: IPPP GOP1 (a), IPPP GOP2 (b), IBBP GOP1 (c), and IBBP GOP2 (d)

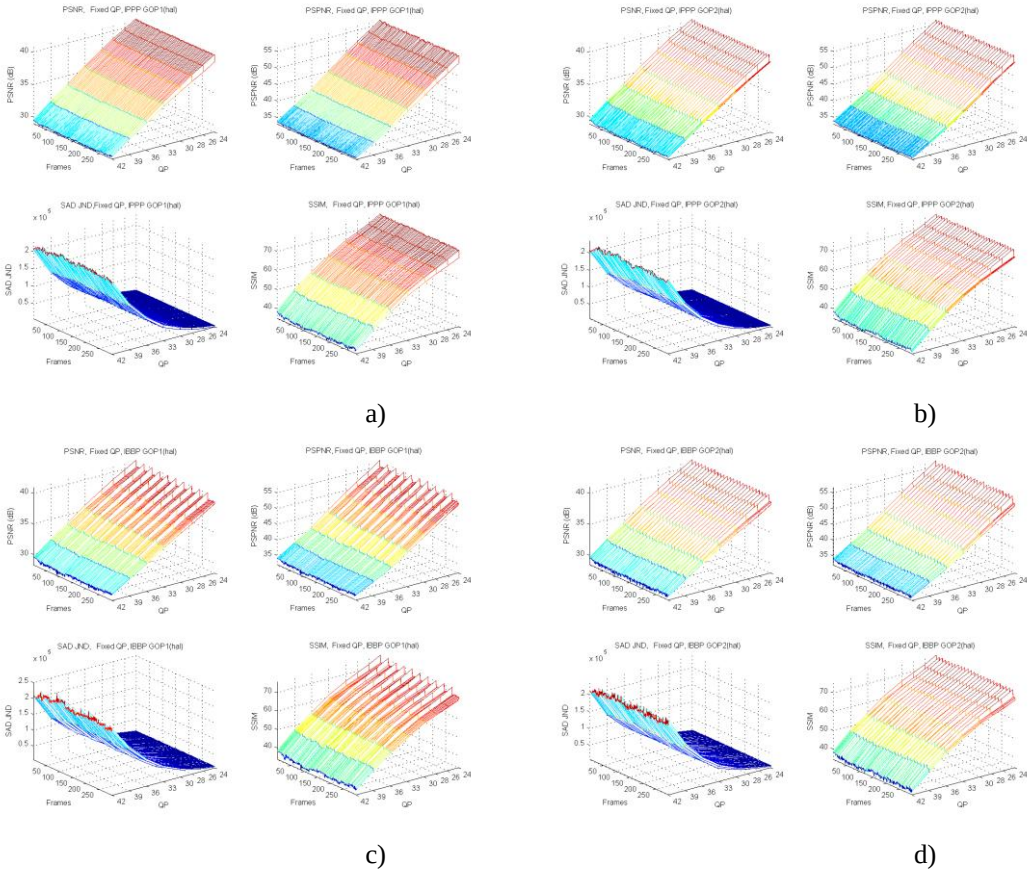


Figure A.9 – Picture quality metrics (PSNR, PSPNR, SAD JND and SSIM) as a function of quantisation for video H.264 video sequence Hall with different GOP structures: IPPP GOP1 (a), IPPP GOP2 (b), IBBP GOP1 (c), and IBBP GOP2 (d)

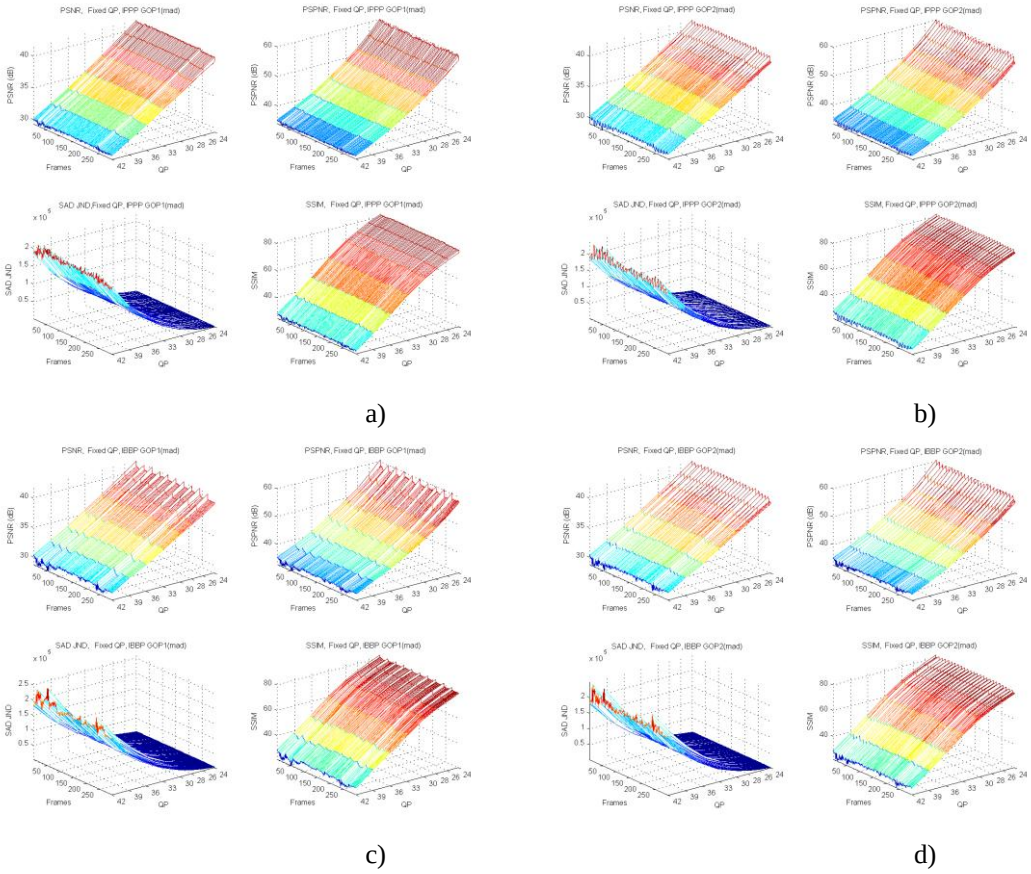


Figure A.10 – Picture quality metrics (PSNR, PSPNR, SAD JND and SSIM) as a function of quantisation for video H.264 video sequence Mother and Daughter with different GOP structures: IPPP GOP1 (a), IPPP GOP2 (b), IBBP GOP1 (c), and IBBP GOP2 (d)

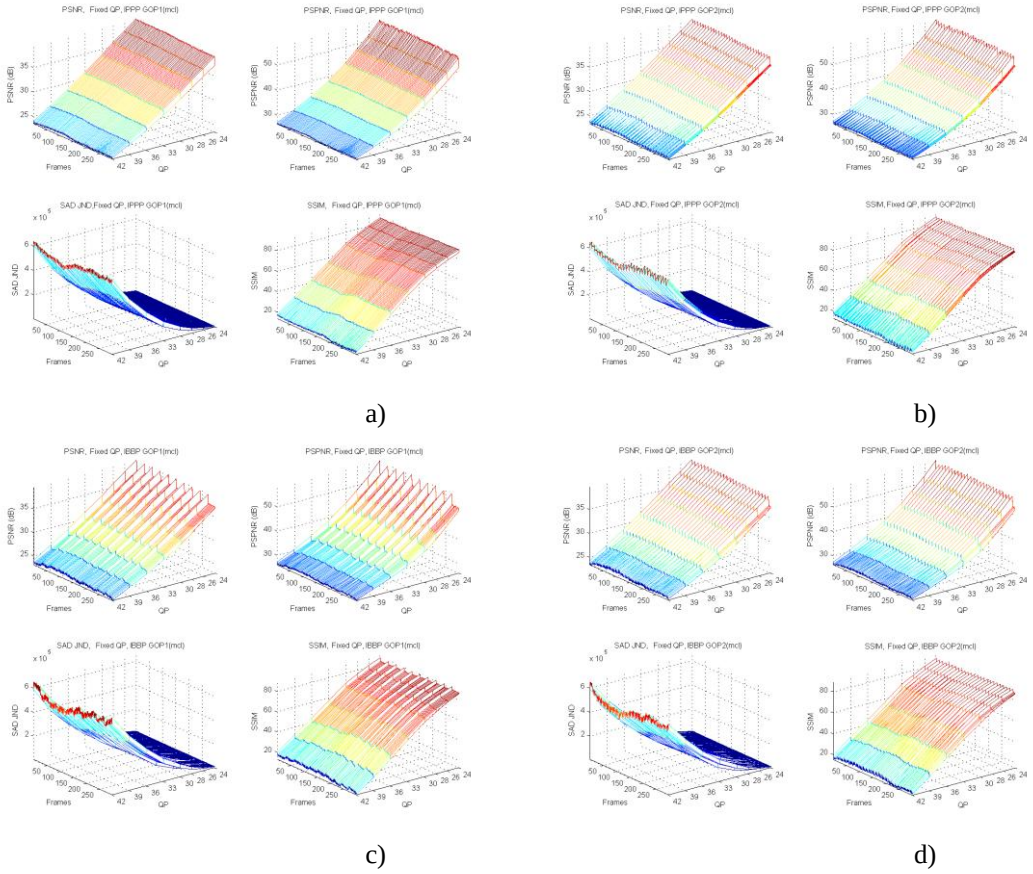


Figure A.11 – Picture quality metrics (PSNR, PSPNR, SAD JND and SSIM) as a function of quantisation for video H.264 video sequence Mobile and Calendar with different GOP structures: IPPP GOP1 (a), IPPP GOP2 (b), IBBP GOP1 (c), and IBBP GOP2 (d)

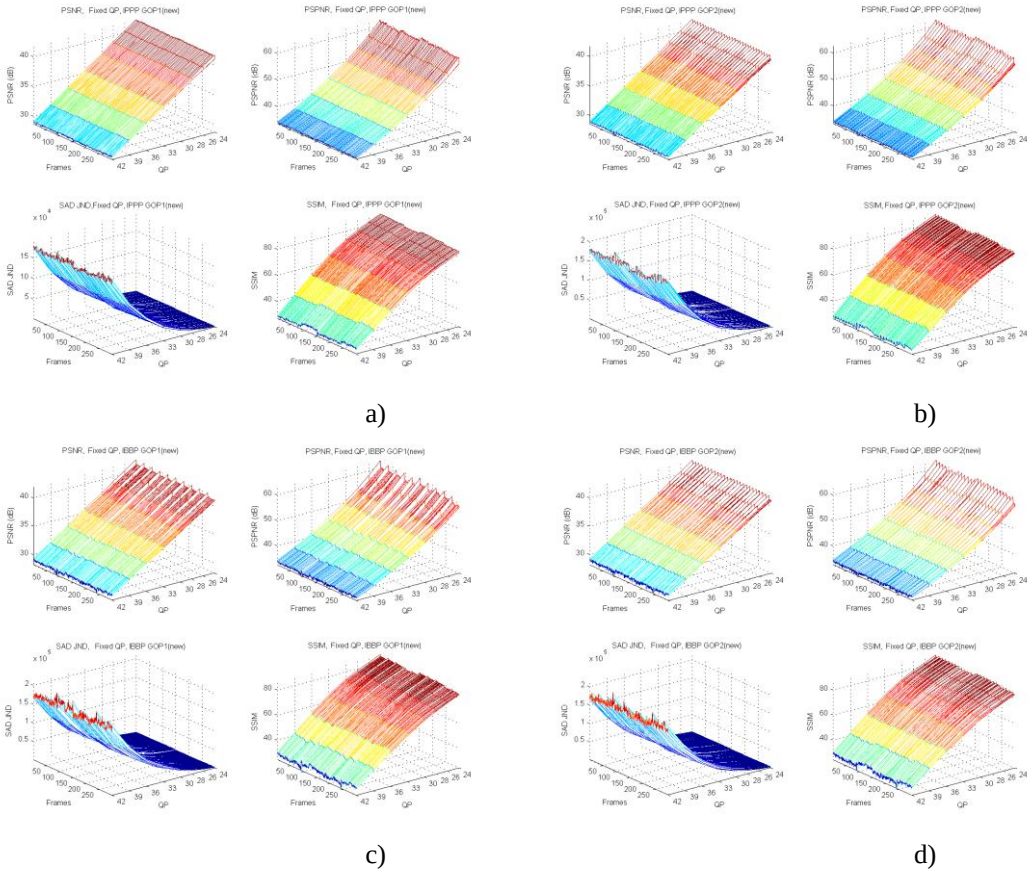


Figure A.12 – Picture quality metrics (PSNR, PSPNR, SAD JND and SSIM) as a function of quantisation for video H.264 video sequence News with different GOP structures: IPPP GOP1 (a), IPPP GOP2 (b), IBBP GOP1 (c), and IBBP GOP2 (d)

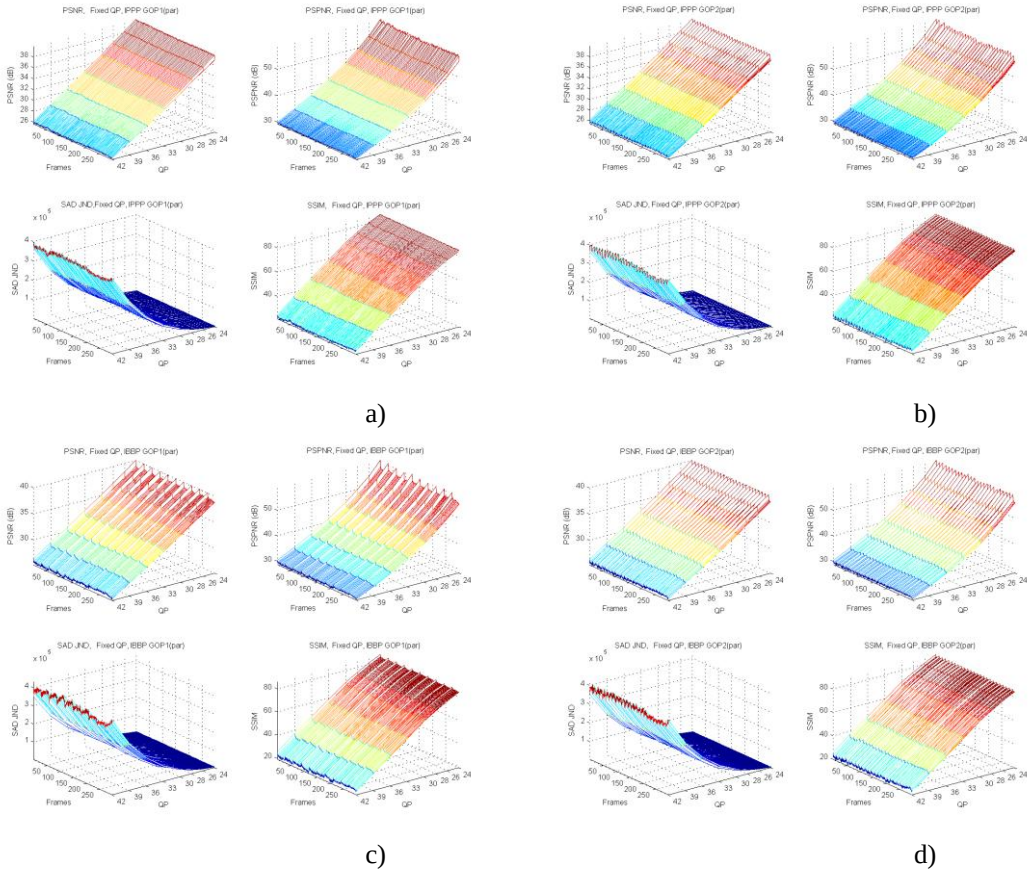


Figure A.13 – Picture quality metrics (PSNR, PSPNR, SAD JND and SSIM) as a function of quantisation for video H.264 video sequence Paris with different GOP structures: IPPP GOP1 (a), IPPP GOP2 (b), IBBP GOP1 (c), and IBBP GOP2 (d)

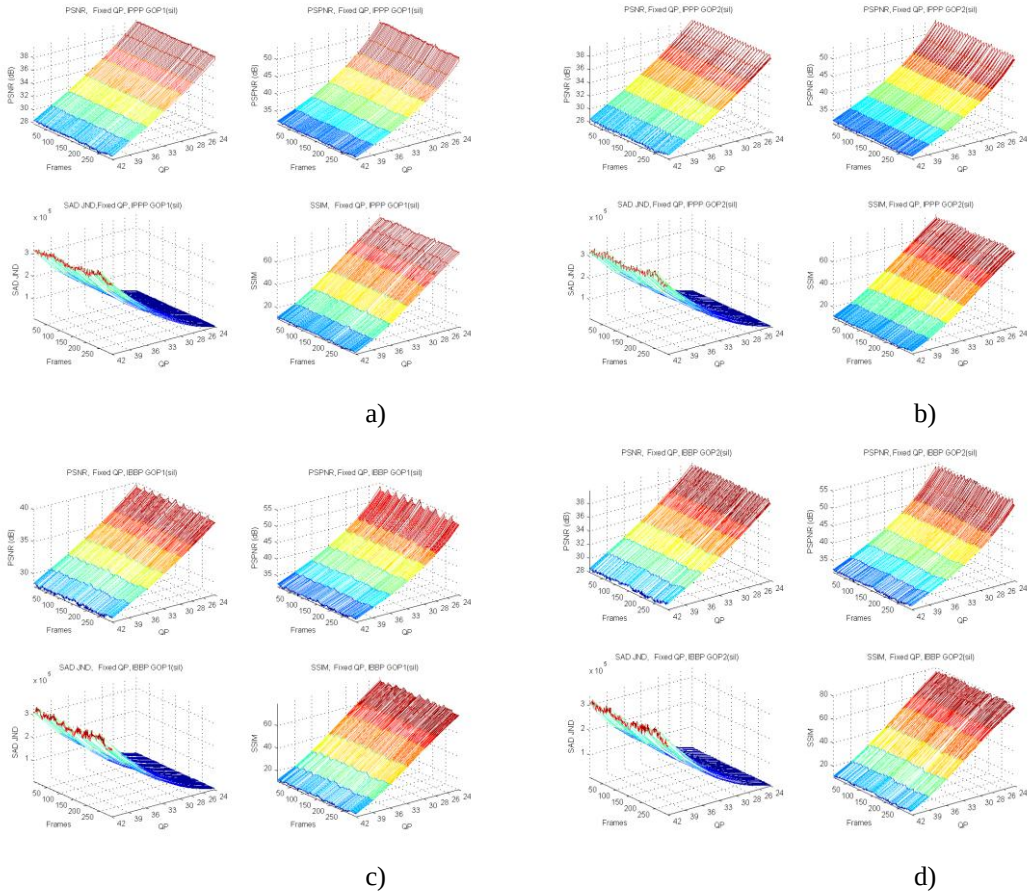


Figure A.14 – Picture quality metrics (PSNR, PSPNR, SAD JND and SSIM) as a function of quantisation for video H.264 video sequence Silence with different GOP structures: IPPP GOP1 (a), IPPP GOP2 (b), IBBP GOP1 (c), and IBBP GOP2 (d)

Annex B. Curve Fitting Data

B.1 Rate-QP and PSNR-QP Curve Fitting Tables (PSNR)

Sequence	Fit Method	Rate – QP		PSNR - QP	
		I Type	P Type	I Type	P Type
AKI	Linear fit	316138390	8142380519	3.8	11.2
	Quadratic fit	23581411	188549137	1.1	3.1
	Exponential fit	35306889	103151127	28.8	87.8
	Logarithmic fit	192484905	3559204542	72.0	218.2
	Power Regression	46364825	2189559228	149.3	456.2
	LNP fit	609908216	36133794248	3167.0	9652.9
CGD	Linear fit	26008897898	112462062075	43.2	119.2
	Quadratic fit	585359589	1896756313	1.3	2.7
	Exponential fit	8043865683	27720522656	5.0	12.0
	Logarithmic fit	15757840576	59020728044	3.9	9.2
	Power Regression	29400473278	113810263889	36.8	110.4
	LNP fit	35294250611	338070552302	2924.1	8726.8
DEA	Linear fit	1771679472	34412261174	2.2	8.1
	Quadratic fit	75884267	318443433	1.1	2.4
	Exponential fit	105112426	5453291471	20.1	57.0
	Logarithmic fit	1006745465	12792358396	48.9	139.6
	Power Regression	613168514	30541753756	153.6	460.0
	LNP fit	4375087816	225292411418	3883.0	11927.7
FLG	Linear fit	40686289539	147408210693	32.7	78.1
	Quadratic fit	832587842	2837470712	2.7	6.3
	Exponential fit	14523704076	77813397217	6.1	28.6
	Logarithmic fit	24409027591	69865661346	15.9	62.5
	Power Regression	48078626034	251822397776	117.3	414.4
	LNP fit	64792014449	953870789507	4374.2	13753.0
FOR	Linear fit	6219640621	57111540802	13.0	42.9
	Quadratic fit	466558379	2162484746	1.3	3.6
	Exponential fit	575823717	9933072106	12.4	33.7
	Logarithmic fit	4213737991	30019379112	33.0	91.7
	Power Regression	564203111	34053053450	85.2	246.1
	LNP fit	4861392720	172626139988	2772.2	8408.7
FOT	Linear fit	19542599661	95609567725	26.3	98.2
	Quadratic fit	590433976	1852386625	1.5	5.2
	Exponential fit	873581083	6996706647	8.2	27.8
	Logarithmic fit	10553059689	47417633966	16.0	40.0
	Power Regression	7979314692	49982622328	53.3	131.4
	LNP fit	45086335695	303468828812	2621.3	7629.5

Table B.1 – Cumulative squared error for Rate-QP and Rate-PSNR curve fitting for video sequences Akiyo, Coastguard, Deadline, Flower Garden, Foreman and Football (IPPP GOP1)

Sequence	Fit Method	Rate – QP		PSNR - QP	
		I Type	P Type	I Type	P Type
HAL	Linear fit	6421227565	33091240461	13.9	42.9
	Quadratic fit	877753324	2906884675	1.0	3.1
	Exponential fit	801758795	2375481769	53.2	169.3
	Logarithmic fit	4745931609	19794960118	103.4	322.1
	Power Regression	131149862	3434261289	190.8	602.8
	LNP fit	2265366676	70287684059	3062.1	9570.6
MAD	Linear fit	1320578714	19508155689	2.3	12.9
	Quadratic fit	110359104	565468248	1.1	3.4
	Exponential fit	120757330	1095749001	10.1	27.0
	Logarithmic fit	837711664	9192562196	37.5	102.9
	Power Regression	248877986	9072091130	96.2	275.1
	LNP fit	1791418338	71973233586	2857.5	8658.3
NEW	Linear fit	851274117	16232750495	1.8	5.6
	Quadratic fit	46538975	300739416	0.9	3.3
	Exponential fit	28062726	585393380	31.7	90.6
	Logarithmic fit	460795608	6362410631	71.5	206.5
	Power Regression	198335635	7602875943	173.9	516.3
	LNP fit	2244375452	92661493972	3732.8	11331.2
PAR	Linear fit	3457765992	51396522491	3.2	12.2
	Quadratic fit	59890473	349321308	1.3	3.0
	Exponential fit	439301132	9366277576	23.4	62.5
	Logarithmic fit	1768503402	17338883171	49.6	135.1
	Power Regression	2443859060	52165341324	171.7	500.2
	LNP fit	11040992102	404169690073	4195.2	12786.9
SIL	Linear fit	2038463813	68647065039	27.6	96.4
	Quadratic fit	78669438	1249645659	0.7	2.8
	Exponential fit	68016596	2962418320	2.7	12.3
	Logarithmic fit	1138300764	35026233214	2.9	6.6
	Power Regression	675550081	30297776753	29.8	79.0
	LNP fit	4080282984	185860921452	2507.7	7494.3
MCL	Linear fit	70394604395	226247141523	39.3	101.9
	Quadratic fit	990985900	1833042216	3.6	9.3
	Exponential fit	17872180054	141258841502	6.1	29.2
	Logarithmic fit	42502904311	105434188189	14.2	56.8
	Power Regression	72150914710	451556762437	112.8	406.3
	LNP fit	104800616387	1324706646442	4343.7	13808.9

Table B.2 – Cumulative squared error for Rate-QP and Rate-PSNR curve fitting for video sequences Hall, Mother and Daughter, News, Paris, Silence and Mobile and Calendar (IPPP GOP1)

Sequence	Fit Method	Rate – QP		PSNR - QP	
		I Type	P Type	I Type	P Type
AKI	Linear fit	192169644	3720726797	0.9	12.5
	Quadratic fit	10403026	129580964	0.5	4.0
	Exponential fit	11088120	108490879	9.1	99.6
	Logarithmic fit	103582965	1759976877	24.8	251.6
	Power Regression	40311178	889765984	54.4	528.6
	LNP fit	563103677	14491637527	1262.2	11402.3
CGD	Linear fit	10535325650	105349761492	17.1	147.5
	Quadratic fit	224611047	2166411298	0.6	4.2
	Exponential fit	2847437665	27820297441	1.8	15.5
	Logarithmic fit	6265084326	60510610835	1.8	14.2
	Power Regression	11089138906	109041938061	17.2	141.7
	LNP fit	16303110085	210401325247	1230.4	10640.2
DEA	Linear fit	939596443	16753588643	1.4	10.3
	Quadratic fit	31019582	304190075	0.6	4.3
	Exponential fit	84242232	2170918610	7.4	67.9
	Logarithmic fit	472336026	7105579361	18.4	167.7
	Power Regression	497423293	12334202079	59.6	540.6
	LNP fit	3581713188	91234347923	1555.6	13962.7
FLG	Linear fit	16361896120	156306280442	11.1	107.5
	Quadratic fit	352702345	3067591789	1.3	10.9
	Exponential fit	5051133172	67196109447	4.8	31.5
	Logarithmic fit	9528185131	86116330836	9.5	70.6
	Power Regression	17725726251	218131815405	60.3	475.1
	LNP fit	33086984332	519204525984	1896.9	16243.2
FOR	Linear fit	2839848317	36648355357	7.0	47.0
	Quadratic fit	190396327	2037148116	0.5	4.6
	Exponential fit	169627973	5478786700	4.7	38.9
	Logarithmic fit	1876654641	21571206873	11.7	108.5
	Power Regression	264815307	14418460056	31.2	293.9
	LNP fit	2621052629	77840831984	1114.6	10006.0
FOT	Linear fit	8335921614	81280376983	11.5	105.2
	Quadratic fit	255106913	2276738488	0.8	5.9
	Exponential fit	397366410	4471263321	4.1	31.3
	Logarithmic fit	4404729120	42709623827	7.6	52.2
	Power Regression	3507001058	36001404207	23.1	174.6
	LNP fit	21361305694	211378169935	1071.5	9249.4

Table B.3 – Cumulative squared error for Rate-QP and Rate-PSNR curve fitting for video sequences Akiyo, Coastguard, Deadline, Flower Garden, Foreman and Football (IPPP GOP2)

Sequence	Fit Method	Rate – QP		PSNR - QP	
		I Type	P Type	I Type	P Type
HAL	Linear fit	2358018150	27184843441	4.7	46.9
	Quadratic fit	322857807	3145676002	0.3	3.6
	Exponential fit	271897667	2841343408	20.2	188.7
	Logarithmic fit	1699237974	18382566679	39.9	366.8
	Power Regression	66140036	1605104769	75.6	686.6
	LNP fit	1344864677	30937452740	1252.5	11136.9
MAD	Linear fit	659628775	10371830098	2.1	13.1
	Quadratic fit	36447899	487301824	0.5	3.5
	Exponential fit	40473489	683025696	2.3	28.8
	Logarithmic fit	380940213	5436830032	10.7	116.4
	Power Regression	191561596	3889715338	32.2	317.3
	LNP fit	1361167558	30388582414	1141.4	10192.2
NEW	Linear fit	521508547	7848486467	0.5	6.2
	Quadratic fit	21786838	220684714	0.5	3.7
	Exponential fit	14678250	288224825	9.8	106.4
	Logarithmic fit	258963985	3385732466	24.0	243.7
	Power Regression	161790554	3228968315	62.7	605.9
	LNP fit	1808622953	38608466589	1479.2	13336.1
PAR	Linear fit	1692059648	27235558327	1.4	12.5
	Quadratic fit	27924709	281539732	0.5	4.4
	Exponential fit	240079965	4508012203	9.6	81.9
	Logarithmic fit	784162075	10685240348	19.8	173.1
	Power Regression	1334634313	25021076378	69.9	612.0
	LNP fit	7539050741	173766852885	1702.2	15115.5
SIL	Linear fit	1415923951	29661882353	11.8	104.7
	Quadratic fit	43226318	652918411	0.3	2.8
	Exponential fit	55897715	1216298944	1.2	11.3
	Logarithmic fit	765714259	15374485429	1.0	9.7
	Power Regression	546285637	12453856019	11.6	103.0
	LNP fit	3185015382	76895389211	1010.9	8977.8
MCL	Linear fit	27378965768	256290267717	14.4	135.0
	Quadratic fit	294490739	2700907131	1.7	14.0
	Exponential fit	8398297109	104459306617	3.6	28.9
	Logarithmic fit	15963785772	141961110090	7.4	60.5
	Power Regression	32125369757	364234389419	53.0	444.4
	LNP fit	52671567659	755529014877	1825.8	16010.3

Table B.4 – Cumulative squared error for Rate-QP and Rate-PSNR curve fitting for video sequences Hall, Mother and Daughter, News, Paris, Silence and Mobile and Calendar (IPPP GOP2)

Sequence	Fit Method	Rate – QP			PSNR - QP		
		I Type	P Type	B Type	I Type	P Type	B Type
AKI	Linear fit	113032240	104629732	1914924637	0.3	3.5	7.3
	Quadratic fit	3593926	16937450	102178656	0.2	1.9	3.7
	Exponential fit	4752348	24818651	77619573	2.7	27.2	61.7
	Logarithmic fit	49978276	78211553	1037341081	7.8	73.9	166.7
	Power Regression	36557508	10681821	358367067	17.3	159.8	362.2
	LNP fit	514718212	33090819	5323564632	414.8	3698.4	8342.9
CGD	Linear fit	4979130334	38714565021	128752998313	10.7	99.8	185.9
	Quadratic fit	331874792	3400345940	3199734359	0.5	5.6	8.1
	Exponential fit	698348940	3193262703	55551309712	2.9	31.6	50.5
	Logarithmic fit	3392787603	28816352492	78964716182	1.1	12.8	28.5
	Power Regression	2976302093	14790633756	190085897798	2.5	19.1	83.1
	LNP fit	5295150656	9661894571	191421754384	362.8	3021.0	7553.9
DEA	Linear fit	560215597	1426912364	9446689408	0.8	4.9	9.3
	Quadratic fit	16249573	121270527	329785989	0.1	1.6	3.6
	Exponential fit	84996655	67693662	823827456	1.6	18.7	47.7
	Logarithmic fit	243015563	998673817	4817502972	4.7	49.5	120.2
	Power Regression	434189930	222186146	4927821833	17.0	165.0	389.0
	LNP fit	3121432743	737406384	34327682994	503.5	4508.6	10213.4
FLG	Linear fit	6740930781	60664726335	138620763081	7.5	80.3	101.9
	Quadratic fit	461717991	4726023574	5956695287	0.5	5.0	9.7
	Exponential fit	1577323987	9394908159	40950535309	1.0	9.8	32.1
	Logarithmic fit	4597261566	44422858660	84017111245	1.5	9.6	55.2
	Power Regression	4905799007	24581901198	126023712402	12.9	81.5	322.0
	LNP fit	12226028090	19435108651	299844228074	584.6	4756.0	11557.0
FOR	Linear fit	748436144	4173102291	23057174504	1.9	15.6	41.1
	Quadratic fit	80004930	768604419	1715618486	0.4	3.7	5.9
	Exponential fit	126094031	1197363808	3213167124	1.3	18.8	30.6
	Logarithmic fit	472892722	3181939681	14341593277	3.5	45.3	78.8
	Power Regression	128131344	558246396	7183000691	10.0	114.6	211.7
	LNP fit	1153163599	1360759232	36874278650	368.9	3439.9	7367.3
FOT	Linear fit	3473243593	19365082161	73335352238	4.5	35.9	89.0
	Quadratic fit	107694492	661571010	2269428173	0.5	5.0	11.1
	Exponential fit	316632732	2466360191	5329018828	1.6	15.9	33.9
	Logarithmic fit	1815270596	11017768673	38525449853	2.2	27.7	50.0
	Power Regression	1945602221	12447561810	36293570052	7.8	87.0	157.5
	LNP fit	9825471612	37038061298	190954848452	379.0	3518.6	7372.0

Table B.5 – Cumulative squared error for Rate-QP and Rate-PSNR curve fitting for video sequences Akiyo, Coastguard, Deadline, Flower Garden, Foreman and Football (IBBP GOP1)

Sequence	Fit Method	Rate – QP			PSNR - QP		
		I Type	P Type	B Type	I Type	P Type	B Type
HAL	Linear fit	677109131	5585474140	17631101454	2.1	16.8	40.1
	Quadratic fit	115783875	1322944283	2493496336	0.2	3.8	5.0
	Exponential fit	99122711	1283259122	1976088707	7.3	56.7	142.4
	Logarithmic fit	466652996	4506524706	12391322799	14.0	111.3	273.8
	Power Regression	58623441	407711636	855462763	25.4	202.7	499.9
	LNP fit	878796023	498504045	13619435587	402.8	3401.8	7997.8
MAD	Linear fit	340158053	545284355	6629301761	1.2	8.3	15.5
	Quadratic fit	13743938	65159804	389426501	0.5	4.0	6.5
	Exponential fit	30270119	74705572	375254564	1.3	11.2	23.5
	Logarithmic fit	170137176	386819300	3825858788	4.2	39.5	87.2
	Power Regression	185686392	185019397	2091907713	11.7	104.4	238.5
	LNP fit	1127171543	288426484	13566238445	392.2	3360.4	7703.7
NEW	Linear fit	261369462	472504774	4840804918	0.4	3.1	5.2
	Quadratic fit	7309210	29735877	195361161	0.3	2.0	3.3
	Exponential fit	20037800	57982789	151043136	4.2	37.4	75.6
	Logarithmic fit	106709873	277182366	2395382896	9.3	83.7	175.2
	Power Regression	153218852	240370296	1597936042	23.5	205.3	439.6
	LNP fit	1411303474	811251578	16753483131	514.7	4448.2	9861.9
PAR	Linear fit	1004795745	3578055517	17150742261	1.1	8.2	11.1
	Quadratic fit	20973072	134414417	264318473	0.2	1.9	3.1
	Exponential fit	175051531	773162649	3167666529	2.0	20.3	56.3
	Logarithmic fit	440743212	2244865014	7899596597	4.8	46.2	120.9
	Power Regression	934723439	3116045101	15458316766	19.7	182.1	440.3
	LNP fit	5628858163	3940464802	76161179039	553.6	4948.9	11145.1
SIL	Linear fit	1126547378	782231723	11929046770	3.6	33.0	75.8
	Quadratic fit	26209752	57550554	313547274	0.1	0.8	1.6
	Exponential fit	59668167	29000951	560538200	0.3	3.1	7.5
	Logarithmic fit	592141787	509524602	6274057124	0.6	4.6	8.3
	Power Regression	525448726	191283641	5123427791	4.8	39.7	80.8
	LNP fit	2812454796	707394231	29829272435	353.6	3111.6	6782.3
MCL	Linear fit	13375089560	117216770544	224798507539	17.9	163.6	214.9
	Quadratic fit	806433238	7260602811	9412518531	0.8	8.4	14.5
	Exponential fit	1817350638	10425949187	32542138313	4.3	36.9	47.7
	Logarithmic fit	9235259397	84831695299	144286428926	2.7	20.7	44.6
	Power Regression	9318545454	61301260242	149807979864	5.5	30.7	195.9
	LNP fit	18303407248	39620749419	345521972654	505.2	4232.0	10424.5

Table B.6 – Cumulative squared error for Rate-QP and Rate-PSNR curve fitting for video sequences Hall, Mother and Daughter, News, Paris, Silence and Mobile and Calendar (IBBP GOP1)

Sequence	Fit Method	Rate – QP			PSNR - QP		
		I Type	P Type	B Type	I Type	P Type	B Type
AKI	Linear fit	136240990	78555139	3437312704	1.2	4.6	10.1
	Quadratic fit	7368257	12965770	130867510	0.5	1.5	3.5
	Exponential fit	8711222	18908914	88607756	8.6	29.7	75.0
	Logarithmic fit	67060809	58823052	1667425916	22.3	73.0	188.3
	Power Regression	37993772	7861339	794104666	47.1	149.0	393.0
	LNP fit	523492100	24634651	12650595166	1042.6	3114.4	8433.1
CGD	Linear fit	11377243024	31152037638	133407341729	28.0	79.2	189.0
	Quadratic fit	829229541	2640488365	3157455160	1.4	5.4	8.4
	Exponential fit	1565234180	2445052735	55273742228	8.4	26.0	53.1
	Logarithmic fit	8104162297	23151687512	80064594984	3.2	10.8	29.2
	Power Regression	6391518251	11065849361	191520910989	5.1	16.5	80.4
	LNP fit	7184914753	7780236919	232198682035	860.6	2444.2	7488.5
DEA	Linear fit	821456262	1174657988	15388256647	1.5	3.1	7.8
	Quadratic fit	37378871	106701973	326126796	0.5	1.1	3.2
	Exponential fit	90804228	57457678	1980508193	5.1	17.6	51.9
	Logarithmic fit	425262134	825224618	6735700891	13.7	44.5	126.5
	Power Regression	471373569	146328541	11149316213	46.0	144.5	405.6
	LNP fit	3253811843	592119767	79892597559	1270.1	3807.6	10407.6
FLG	Linear fit	18078373065	47325384748	138899282945	26.2	61.5	109.9
	Quadratic fit	1203051026	4067581815	6156300376	1.1	5.2	11.4
	Exponential fit	2681089039	7499608572	46646909283	3.2	10.8	31.3
	Logarithmic fit	12828640409	35026916542	81307790133	2.2	11.0	51.2
	Power Regression	8966079354	15725064717	140828415674	19.0	67.0	291.6
	LNP fit	16032945479	13361870465	439675158161	1314.6	3756.8	11174.1
FOR	Linear fit	1519557865	3226492867	29082326660	4.0	10.6	35.7
	Quadratic fit	199510421	598489824	1668858240	1.1	3.2	6.5
	Exponential fit	286217984	1012963527	4576229488	5.5	16.4	34.0
	Logarithmic fit	1051666714	2464665203	16645858387	13.4	40.3	87.3
	Power Regression	184880576	481166427	12980222372	33.1	100.3	228.2
	LNP fit	1439142792	1015322239	67825065486	961.2	2863.5	7430.8
FOT	Linear fit	6510684103	16087975645	79983028513	9.3	30.0	88.8
	Quadratic fit	186827659	588990082	2202595507	1.1	5.2	10.3
	Exponential fit	684726957	2032935442	6205411744	3.6	12.8	32.9
	Logarithmic fit	3506026832	9188084500	41125708510	7.0	22.0	49.3
	Power Regression	3936921021	10090735155	41658953963	24.0	69.6	157.4
	LNP fit	15932499070	30300847172	225492074753	1000.8	2881.2	7425.7

Table B.7 – Cumulative squared error for Rate-QP and Rate-PSNR curve fitting for video sequences Akiyo, Coastguard, Deadline, Flower Garden, Foreman and Football (IBBP GOP2)

Sequence	Fit Method	Rate – QP			PSNR - QP		
		I Type	P Type	B Type	I Type	P Type	B Type
HAL	Linear fit	1614408488	5524144188	21506066582	4.9	16.6	44.6
	Quadratic fit	341559506	1340482984	2782235894	1.0	4.3	6.4
	Exponential fit	340289631	1293448794	2212494156	16.2	50.8	150.4
	Logarithmic fit	1225712833	4477489375	14318173000	31.7	96.8	284.0
	Power Regression	143942615	390793739	1556375931	57.4	172.6	513.7
	LNP fit	950111415	445530602	27278768806	953.5	2803.4	8047.7
MAD	Linear fit	432117391	440212704	9759894163	2.1	4.4	11.4
	Quadratic fit	25962183	50613061	427065435	1.0	3.3	6.2
	Exponential fit	42236700	52729916	629256645	3.2	13.6	32.1
	Logarithmic fit	237649108	309075661	5126371865	11.3	42.7	106.2
	Power Regression	208953252	138541192	4050179654	29.9	103.5	270.4
	LNP fit	1158921583	253161893	27997562207	952.9	2865.1	7849.6
NEW	Linear fit	329427955	403171226	7632317254	0.8	2.9	5.4
	Quadratic fit	11797137	27209924	223169605	0.5	1.6	3.3
	Exponential fit	23600766	53368863	257807332	10.1	34.2	78.7
	Logarithmic fit	148973192	235217594	3392978104	23.0	74.4	180.2
	Power Regression	173729899	210748080	3014173722	56.8	178.9	449.0
	LNP fit	1500778146	709934947	34753295891	1244.4	3742.4	9934.0
PAR	Linear fit	1608322721	2895592580	25459747209	2.4	6.1	11.1
	Quadratic fit	39559464	106115675	273813617	0.6	1.5	3.3
	Exponential fit	331667669	568940040	4752975763	5.7	17.8	58.2
	Logarithmic fit	814745105	1813046598	10223231375	13.0	40.0	123.7
	Power Regression	1511680504	2382765926	24633942562	51.1	154.4	447.9
	LNP fit	6349194185	3231981022	155163737991	1391.4	4133.8	11266.5
SIL	Linear fit	1204152780	693871717	25300180634	9.0	28.5	79.0
	Quadratic fit	33330147	48789704	515556464	0.3	0.8	2.2
	Exponential fit	62329821	23891082	1289764811	0.9	2.7	8.5
	Logarithmic fit	645759367	448756199	12945281096	1.3	3.8	7.9
	Power Regression	530550537	174262137	11706796033	10.9	33.9	79.9
	LNP fit	2856888765	640296716	68621270275	857.5	2656.5	6859.1
MCL	Linear fit	32722261895	93941886717	221549376107	50.5	126.0	217.0
	Quadratic fit	2017151324	6067856492	9189069911	2.4	8.3	15.5
	Exponential fit	3471442822	7771635282	36444777828	13.5	31.6	52.7
	Logarithmic fit	23240896166	68150684595	136790433359	8.0	18.6	45.9
	Power Regression	18805769088	46033975604	162382588425	8.7	30.9	191.3
	LNP fit	24608454822	30781005499	536476772610	1164.4	3430.4	10267.9

Table B.8 – Cumulative squared error for Rate-QP and Rate-PSNR curve fitting for video sequences Hall, Mother and Daughter, News, Paris, Silence and Mobile and Calendar (IBBP GOP2)

B.2 Rate-PSNR Curve Fitting Tables (PSNR)

Sequence	Fit Method	IPPP GOP1		IPPP GOP2	
		I Type	P Type	I Type	P Type
AKI	Linear fit	339704889	8796244965	205988745	4017395398
	Quadratic fit	30200413	306626785	13508910	182848495
	Exponential fit	43482700	114544393	11804693	126048302
	Logarithmic fit	420962419	12109825080	267230256	5415213718
	Power Regression	78696839	680375741	28488902	418510037
	LNP fit	2484099855	112237484165	1973201334	46512410837
CGD	Linear fit	16821012235	65746896716	6743059150	66429490022
	Quadratic fit	117193593	633721831	56277245	536848578
	Exponential fit	22613301866	94195503847	8508089006	84284115494
	Logarithmic fit	23238063112	100915928085	9474686760	95577933914
	Power Regression	10199129456	39562020848	3627693746	35698377317
	LNP fit	107855874746	676287171870	46264759132	517328302919
DEA	Linear fit	1632104886	28590764159	832741314	14359971604
	Quadratic fit	75188585	265935643	30546751	288646286
	Exponential fit	158272463	8977200277	140926246	3493415020
	Logarithmic fit	2251184255	48242742633	1231492270	22930193593
	Power Regression	122655605	1123433351	49129789	679992156
	LNP fit	10545259017	372369622558	7089795820	159132562923
FLG	Linear fit	30373569572	99488794398	12678235586	113472375671
	Quadratic fit	373597766	1619947645	201703664	1636483893
	Exponential fit	25151836031	136582659756	7965178400	113275458375
	Logarithmic fit	43962923259	176858382831	18783620069	177285707065
	Power Regression	8139024849	42796748445	2087522116	35862609072
	LNP fit	201650364694	1508989890230	91469676274	1076705086732
FOR	Linear fit	5251843198	39927993567	2307745613	27714527403
	Quadratic fit	275753138	1111865314	97678322	1115655386
	Exponential fit	529781112	25926483849	200504240	11126514699
	Logarithmic fit	6505782289	57664290310	2914541252	37457933965
	Power Regression	744436520	12870353134	215486585	6825676374
	LNP fit	27861161359	447065222469	13553532520	235264092549
FOT	Linear fit	13395214258	51669093589	5619927704	51089191747
	Quadratic fit	279646152	1060869810	133668955	1131768650
	Exponential fit	5204341376	50566771436	2396511190	29461669273
	Logarithmic fit	19014446862	80138936717	8120245089	74502890699
	Power Regression	1664700663	20349465141	752545971	10861940711
	LNP fit	110603896960	676103551212	49643075788	507949955743

Table B.9 – Cumulative squared error for Rate-PSNR curve fitting (IPPP GOP1; IPPP GOP2)

Sequence	Fit Method	IPPP GOP1		IPPP GOP2	
		I Type	P Type	I Type	P Type
HAL	Linear fit	7338039622	41505140430	2714180289	32267472030
	Quadratic fit	1230960499	4604365754	459295404	4587565174
	Exponential fit	1529576368	5383202060	521565942	5569915247
	Logarithmic fit	8341355285	51562382447	3127572762	38317888005
	Power Regression	2552830831	11065995223	896030728	10035317565
	LNP fit	24697031523	279113554689	10456072523	160108808824
MAD	Linear fit	1187977024	15117907823	535966417	8513322084
	Quadratic fit	104338691	422689579	28673827	401781344
	Exponential fit	153852828	3005308414	78566432	1355072085
	Logarithmic fit	1471466869	21645691683	702785209	11612240904
	Power Regression	185798863	723304835	47147288	639384917
	LNP fit	8479448117	232048726026	5347042290	105294046787
NEW	Linear fit	878195916	16020862383	520472718	7861691078
	Quadratic fit	54277517	412597268	23517162	274455774
	Exponential fit	32451604	592319534	15915322	283325520
	Logarithmic fit	1178843849	24311362109	729856469	11553484311
	Power Regression	82420317	471419054	28390828	361718809
	LNP fit	8426206244	269295147913	6136876567	116604799377
PAR	Linear fit	3117039910	40910125593	1486163009	22566475887
	Quadratic fit	56330220	215166259	25728655	229896302
	Exponential fit	584473602	16767981217	350556152	7542281094
	Logarithmic fit	4615040641	74970918367	2332212969	38625281569
	Power Regression	61375936	1460604684	32813272	689862418
	LNP fit	24171836202	634450446064	14277201676	289750789941
SIL	Linear fit	1213808369	36401430003	789887463	16205638640
	Quadratic fit	24001814	131212631	11619480	110721586
	Exponential fit	631308838	29598098934	549142172	11858095729
	Logarithmic fit	1715594918	55370554570	1148923518	24304878472
	Power Regression	223993903	10525823324	204903952	4140230653
	LNP fit	9543226722	527032201207	6973033268	215686829934
MCL	Linear fit	51001973321	148642109387	20075738208	180105744654
	Quadratic fit	329486337	888663438	124010704	1067888412
	Exponential fit	37191641431	247326781254	15519298941	191692405346
	Logarithmic fit	74437525279	268812940150	30112387806	283370752245
	Power Regression	9338080607	79263965020	3703193443	55959092712
	LNP fit	338854804146	2201791734017	147657802214	1659689677857

Table B.10 – Cumulative squared error for Rate-PSNR curve fitting (IPPP GOP1; IPPP GOP2)

Sequence	Fit Method	IBBP GOP1			IBBP GOP2		
		I Type	P Type	B Type	I Type	P Type	B Type
AKI	Linear fit	1,26E+08	1,08E+08	2,04E+09	1,50E+08	8,25E+07	3,71E+09
	Quadratic fit	6,20E+06	1,94E+07	1,33E+08	1,04E+07	1,48E+07	1,75E+08
	Exponential fit	4,75E+06	2,80E+07	9,52E+07	9,18E+06	2,22E+07	1,08E+08
	Logarithmic fit	1,73E+08	1,22E+08	2,64E+09	2,01E+08	9,35E+07	4,97E+09
	Power Regression	1,29E+07	4,15E+07	2,52E+08	2,01E+07	3,27E+07	3,65E+08
	LNP fit	1,60E+09	3,80E+08	1,90E+10	1,69E+09	2,84E+08	4,14E+10
CGD	Linear fit	3,01E+09	2,52E+10	7,60E+10	7,13E+09	2,03E+10	7,70E+10
	Quadratic fit	7,74E+07	6,67E+08	4,66E+08	1,60E+08	4,76E+08	4,98E+08
	Exponential fit	3,75E+09	2,20E+10	1,84E+11	8,27E+09	1,71E+10	1,87E+11
	Logarithmic fit	3,98E+09	3,13E+10	1,07E+11	9,12E+09	2,51E+10	1,10E+11
	Power Regression	1,59E+09	8,49E+09	9,36E+10	3,49E+09	6,66E+09	9,53E+10
	LNP fit	1,71E+10	9,41E+10	5,30E+11	3,31E+10	7,60E+10	5,89E+11
DEA	Linear fit	4,65E+08	1,35E+09	8,49E+09	7,10E+08	1,12E+09	1,33E+10
	Quadratic fit	1,30E+07	1,14E+08	3,25E+08	3,27E+07	1,01E+08	3,18E+08
	Exponential fit	1,47E+08	5,52E+07	1,22E+09	1,52E+08	4,64E+07	3,09E+09
	Logarithmic fit	7,48E+08	1,65E+09	1,24E+10	1,05E+09	1,37E+09	2,10E+10
	Power Regression	3,35E+07	1,42E+08	4,42E+08	5,22E+07	1,36E+08	6,41E+08
	LNP fit	5,36E+09	4,73E+09	6,97E+10	6,22E+09	3,88E+09	1,42E+11
FLG	Linear fit	4,91E+09	4,44E+10	1,03E+11	1,26E+10	3,50E+10	9,81E+10
	Quadratic fit	1,75E+08	1,61E+09	2,49E+09	3,40E+08	1,47E+09	2,58E+09
	Exponential fit	3,39E+09	2,12E+10	6,92E+10	7,90E+09	1,54E+10	8,28E+10
	Logarithmic fit	6,84E+09	5,72E+10	1,51E+11	1,69E+10	4,44E+10	1,49E+11
	Power Regression	1,30E+09	9,89E+09	2,50E+10	2,92E+09	8,45E+09	3,06E+10
	LNP fit	3,12E+10	1,75E+11	7,56E+11	6,35E+10	1,31E+11	8,91E+11
FOR	Linear fit	6,58E+08	3,77E+09	1,83E+10	1,36E+09	2,94E+09	2,20E+10
	Quadratic fit	4,58E+07	4,65E+08	9,58E+08	1,15E+08	3,76E+08	9,35E+08
	Exponential fit	7,83E+07	5,84E+08	5,04E+09	1,39E+08	5,61E+08	9,37E+09
	Logarithmic fit	8,36E+08	4,40E+09	2,38E+10	1,66E+09	3,42E+09	3,00E+10
	Power Regression	1,27E+08	1,11E+09	3,54E+09	2,75E+08	9,98E+08	5,38E+09
	LNP fit	5,09E+09	1,36E+10	1,34E+11	7,75E+09	1,04E+10	2,02E+11
FOT	Linear fit	2,05E+09	1,46E+10	4,79E+10	4,43E+09	1,22E+10	4,97E+10
	Quadratic fit	5,05E+07	4,48E+08	1,09E+09	1,28E+08	3,85E+08	1,21E+09
	Exponential fit	1,59E+09	6,14E+09	2,60E+10	2,51E+09	4,77E+09	3,32E+10
	Logarithmic fit	3,12E+09	2,04E+10	6,99E+10	6,44E+09	1,69E+10	7,39E+10
	Power Regression	5,35E+08	1,86E+09	9,27E+09	8,16E+08	1,41E+09	1,24E+10
	LNP fit	2,07E+10	9,97E+10	4,53E+11	3,69E+10	8,28E+10	5,27E+11

Table B.11 – Cumulative squared error for Rate-PSNR curve fitting (IBBP GOP1; IBBP GOP2)

Sequence	Fit Method	IBBP GOP1			IBBP GOP2		
		I Type	P Type	B Type	I Type	P Type	B Type
HAL	Linear fit	8,15E+08	5,99E+09	2,06E+10	1,81E+09	5,93E+09	2,59E+10
	Quadratic fit	1,47E+08	1,41E+09	3,26E+09	3,79E+08	1,40E+09	3,65E+09
	Exponential fit	1,64E+08	1,67E+09	3,60E+09	4,57E+08	1,66E+09	4,10E+09
	Logarithmic fit	9,65E+08	6,58E+09	2,40E+10	2,05E+09	6,50E+09	3,09E+10
	Power Regression	2,79E+08	2,51E+09	6,32E+09	7,09E+08	2,51E+09	7,53E+09
	LNP fit	4,08E+09	1,47E+10	8,74E+10	6,50E+09	1,43E+10	1,34E+11
MAD	Linear fit	2,56E+08	5,08E+08	5,80E+09	3,43E+08	4,27E+08	8,36E+09
	Quadratic fit	1,15E+07	6,62E+07	3,65E+08	2,32E+07	5,67E+07	4,00E+08
	Exponential fit	7,39E+07	7,74E+07	5,78E+08	8,83E+07	5,30E+07	1,18E+09
	Logarithmic fit	3,60E+08	5,96E+08	7,52E+09	4,60E+08	5,02E+08	1,13E+10
	Power Regression	2,23E+07	7,78E+07	3,66E+08	3,88E+07	6,08E+07	4,95E+08
	LNP fit	3,67E+09	2,33E+09	5,42E+10	4,01E+09	1,96E+09	9,81E+10
NEW	Linear fit	2,58E+08	4,95E+08	4,88E+09	3,28E+08	4,28E+08	7,64E+09
	Quadratic fit	8,33E+06	3,56E+07	2,33E+08	1,38E+07	3,29E+07	2,84E+08
	Exponential fit	2,06E+07	4,44E+07	1,36E+08	2,30E+07	3,87E+07	2,46E+08
	Logarithmic fit	3,88E+08	6,39E+08	6,78E+09	4,76E+08	5,53E+08	1,11E+10
	Power Regression	6,19E+06	3,62E+07	2,57E+08	1,09E+07	3,26E+07	3,30E+08
	LNP fit	4,14E+09	3,70E+09	5,68E+10	4,60E+09	3,21E+09	1,08E+11
PAR	Linear fit	8,23E+08	3,19E+09	1,48E+10	1,36E+09	2,60E+09	2,11E+10
	Quadratic fit	1,42E+07	9,49E+07	2,11E+08	2,75E+07	7,53E+07	2,03E+08
	Exponential fit	3,18E+08	1,06E+09	4,68E+09	5,20E+08	7,70E+08	7,76E+09
	Logarithmic fit	1,36E+09	4,26E+09	2,33E+10	2,08E+09	3,47E+09	3,57E+10
	Power Regression	3,65E+07	2,44E+08	6,22E+08	9,10E+07	1,64E+08	8,50E+08
	LNP fit	9,72E+09	1,57E+10	1,46E+11	1,24E+10	1,28E+10	2,63E+11
SIL	Linear fit	6,00E+08	5,76E+08	6,95E+09	6,57E+08	5,07E+08	1,39E+10
	Quadratic fit	6,46E+06	2,60E+07	7,89E+07	9,53E+06	2,01E+07	8,94E+07
	Exponential fit	5,45E+08	1,10E+08	4,40E+09	5,50E+08	1,05E+08	1,09E+10
	Logarithmic fit	8,97E+08	7,34E+08	1,02E+10	9,67E+08	6,51E+08	2,10E+10
	Power Regression	2,12E+08	3,13E+07	1,50E+09	2,15E+08	2,81E+07	3,91E+09
	LNP fit	1,01E+10	3,11E+09	9,93E+10	1,04E+10	3,19E+09	2,32E+11
MCL	Linear fit	8,20E+09	7,31E+10	1,49E+11	1,95E+10	5,95E+10	1,42E+11
	Quadratic fit	1,19E+08	1,11E+09	1,87E+09	2,58E+08	9,55E+08	1,78E+09
	Exponential fit	1,04E+10	7,84E+10	1,12E+11	2,46E+10	5,90E+10	1,22E+11
	Logarithmic fit	1,17E+10	9,73E+10	2,15E+11	2,67E+10	7,86E+10	2,14E+11
	Power Regression	3,08E+09	2,39E+10	2,79E+10	7,91E+09	1,82E+10	3,17E+10
	LNP fit	5,40E+10	3,34E+11	1,02E+12	1,07E+11	2,66E+11	1,21E+12

Table B.12 – Cumulative squared error for Rate-PSNR curve fitting (IBBP GOP1)

B.3 Curve Fitting Tables (SAD_JND; SSD_JND; PSPNR)

Sequence	Fit Method	SAD_JND-QP		Rate - SAD_JND	
		I Type	P Type	I Type	P Type
Akiyo	Linear fit	4591	2685	3967	14452
	Quadratic fit	857	493	2485	8532
	Exponential fit	3612	2200	3327	11532
	Logarithmic fit	5560	3243	1312	2460
	Power Regression	1291	827	1222	5283
	LNP fit	11288	6531	6771	27294
Foreman	Linear fit	8179	4775	13980	30381
	Quadratic fit	1213	711	8176	16513
	Exponential fit	9714	6295	11075	21172
	Logarithmic fit	10565	6131	5832	7531
	Power Regression	5394	3694	6678	23278
	LNP fit	25447	14536	23495	62379
Football	Linear fit	8111	4457	27699	37685
	Quadratic fit	794	401	14704	19338
	Exponential fit	14491	8689	17711	23905
	Logarithmic fit	10918	5983	5999	6201
	Power Regression	8844	5507	21006	33678
	LNP fit	28680	15608	57947	82739

Table B.13 – D-QP and R-D Mean Absolute Error (SAD_JND; IPPP GOP1, Akiyo, Foreman, Football)

Sequence	Fit Method	SAD_JND-QP		Rate - SAD_JND	
		I Type	P Type	I Type	P Type
Akiyo	Linear fit	7215	2441	5391	8565
	Quadratic fit	1278	448	3225	5108
	Exponential fit	6150	1986	4374	6880
	Logarithmic fit	8790	2958	1379	1711
	Power Regression	2410	741	1898	3093
	LNP fit	18140	6014	9680	15955
Foreman	Linear fit	12838	4382	15269	20532
	Quadratic fit	1915	637	8722	11467
	Exponential fit	16992	5450	11712	14958
	Logarithmic fit	16741	5658	5911	6401
	Power Regression	9956	3081	7782	14107
	LNP fit	41290	13591	26267	40071
Football	Linear fit	13167	4218	29258	30458
	Quadratic fit	1501	441	15603	16003
	Exponential fit	23774	7810	18917	19411
	Logarithmic fit	17700	5691	6175	6057
	Power Regression	14634	4836	22504	24796
	LNP fit	46467	15031	61786	65077

Table B.14 – D-QP and R-D Mean Absolute Error (SAD_JND; IPPP GOP2, Akiyo, Foreman, Football)

Sequence	Fit Method	SAD_JND-QP			Rate - SAD_JND		
		I Type	P Type	B Type	I Type	P Type	B Type
Akiyo	Linear fit	12111	4075	2761	8274	1624	6466
	Quadratic fit	2126	731	483	4831	1185	3921
	Exponential fit	10807	3508	2370	6645	1488	5214
	Logarithmic fit	14720	4943	3358	1452	837	1802
	Power Regression	4404	1388	936	2966	358	2057
	LNP fit	30170	10080	6897	15265	2280	11719
Foreman	Linear fit	19066	6672	5254	15252	9424	18110
	Quadratic fit	2934	880	798	8923	6399	10162
	Exponential fit	29506	10123	6060	12209	8232	13354
	Logarithmic fit	25157	8789	6797	5490	5022	6752
	Power Regression	18072	6053	3253	5055	2369	10715
	LNP fit	63678	22085	16438	26648	14159	33715
Football	Linear fit	20395	7149	5722	32940	24908	33501
	Quadratic fit	2393	807	843	17019	13344	18081
	Exponential fit	41919	13775	10295	20550	15427	21237
	Logarithmic fit	27656	9608	7692	6612	5723	7143
	Power Regression	27047	8707	6299	31440	22894	25130
	LNP fit	73997	25232	20211	71067	49603	71825

Table B.15 – D-QP and R-D Mean Absolute Error (SAD_JND; IBBP GOP1, Akiyo, Foreman, Football)

Sequence	Fit Method	PSNR-QP			Rate - PSNR		
		I Type	P Type	B Type	I Type	P Type	B Type
Akiyo	Linear fit	7712	4532	2802	5435	1550	9349
	Quadratic fit	1466	871	523	3237	1133	5516
	Exponential fit	6272	3752	2216	4400	1427	7463
	Logarithmic fit	9332	5463	3389	1175	793	2226
	Power Regression	2364	1436	803	1895	324	3023
	LNP fit	18907	10944	6853	9881	2161	17445
Foreman	Linear fit	11939	7394	5260	12378	9132	22193
	Quadratic fit	1512	972	844	7600	6238	12086
	Exponential fit	16552	11014	5450	10209	8078	15993
	Logarithmic fit	15678	9608	6727	5336	4819	7330
	Power Regression	9594	6607	2761	3950	2058	13397
	LNP fit	39089	23426	15829	20610	13580	42883
Football	Linear fit	13680	7688	5704	28448	24960	35502
	Quadratic fit	1597	780	803	15085	13636	19009
	Exponential fit	25752	15871	10277	17963	15725	22500
	Logarithmic fit	18259	10288	7648	6070	5508	7235
	Power Regression	16242	10298	6288	25916	23263	26909
	LNP fit	47299	26722	19957	58754	49410	77004

Table B.16 – D-QP and R-D Mean Absolute Error (SAD_JND; IBBP GOP2, Akiyo, Foreman, Football)

Sequence	Fit Method	SSD_JND-QP		Rate - SSD_JND	
		I Type	P Type	I Type	P Type
Akiyo	Linear fit	84457	48424	4682	17743
	Quadratic fit	26673	15370	3440	12482
	Exponential fit	54143	31993	4354	15987
	Logarithmic fit	95438	54634	1439	3199
	Power Regression	11316	7090	1065	4326
	LNP fit	155595	88552	6771	27294
Foreman	Linear fit	195692	111717	16426	38229
	Quadratic fit	44971	25410	11722	25985
	Exponential fit	206399	132877	15181	32649
	Logarithmic fit	229413	130534	6225	8045
	Power Regression	103483	69635	5651	20453
	LNP fit	422963	237905	23495	62379
Football	Linear fit	231366	120503	36336	50434
	Quadratic fit	50121	26184	24669	33490
	Exponential fit	262700	152103	30770	42404
	Logarithmic fit	270847	140987	8187	8110
	Power Regression	126858	78847	17363	28847
	LNP fit	496574	257967	57947	82739

Table B.17 – D-QP and R-D Mean Absolute Error (SSD_JND; IPPP GOP1, Akiyo, Foreman, Football)

Sequence	Fit Method	SSD_JND-QP		Rate - SSD_JND	
		I Type	P Type	I Type	P Type
Akiyo	Linear fit	137962	45221	6470	10462
	Quadratic fit	42680	14255	4617	7402
	Exponential fit	95853	30006	5903	9465
	Logarithmic fit	156253	51101	1570	2115
	Power Regression	24052	6763	1638	2513
	LNP fit	256796	83311	9680	15955
Foreman	Linear fit	317087	105074	18039	25364
	Quadratic fit	69496	24030	12670	17533
	Exponential fit	373107	114420	16359	22183
	Logarithmic fit	373648	123004	6340	6846
	Power Regression	199940	57623	6604	12295
	LNP fit	700159	225607	26267	40071
Football	Linear fit	380297	120833	38582	40289
	Quadratic fit	82212	26047	26219	27140
	Exponential fit	438797	141329	32774	33997
	Logarithmic fit	445239	141556	8281	8162
	Power Regression	215730	70013	18700	20892
	LNP fit	816931	260187	61786	65077

Table B.18 – D-QP and R-D Mean Absolute Error (SSD_JND; IPPP GOP2, Akiyo, Foreman, Football)

Sequence	Fit Method	SSD_JND-QP			Rate - SSD_JND		
		I Type	P Type	B Type	I Type	P Type	B Type
Akiyo	Linear fit	220074	73249	51602	9982	1816	7822
	Quadratic fit	68968	23029	15884	7008	1476	5554
	Exponential fit	160820	51589	37560	9010	1796	7109
	Logarithmic fit	248943	82818	58471	1824	882	2031
	Power Regression	43462	13281	10593	2406	342	1713
	LNP fit	407388	135302	96259	15265	2280	11719
Foreman	Linear fit	472941	165700	134357	17773	10678	21993
	Quadratic fit	88681	32116	31511	12371	8361	15110
	Exponential fit	674960	219834	130173	15992	10339	19369
	Logarithmic fit	561262	195924	157408	5714	5273	7326
	Power Regression	382945	119450	61947	4922	2086	9550
	LNP fit	1072588	370288	289905	26648	14159	33715
Football	Linear fit	568307	197659	192710	44019	32007	45060
	Quadratic fit	119985	43337	45876	29435	21902	30818
	Exponential fit	756576	247944	206895	37092	27021	38308
	Logarithmic fit	667787	231602	224180	8018	7575	10206
	Power Regression	406202	128879	96113	27210	19391	20233
	LNP fit	1240117	426282	403311	71067	49603	71825

Table B.19 – D-QP and R-D Mean Absolute Error (SSD_JND; IBBP GOP1, Akiyo, Foreman, Football)

Sequence	Fit Method	PSNR-QP			Rate - PSNR		
		I Type	P Type	B Type	I Type	P Type	B Type
Akiyo	Linear fit	138491	79596	51735	6525	1730	11363
	Quadratic fit	44340	25853	16347	4630	1407	8021
	Exponential fit	90994	52871	34085	5920	1715	10249
	Logarithmic fit	156377	89695	58456	1390	839	2604
	Power Regression	20734	11890	8046	1540	307	2504
	LNP fit	254329	144825	95277	9881	2161	17445
Foreman	Linear fit	284502	175224	129649	14290	10316	27380
	Quadratic fit	57556	37047	32879	10341	8105	18512
	Exponential fit	345996	231222	110191	13225	10033	23793
	Logarithmic fit	335974	205717	150875	5567	5049	8155
	Power Regression	182412	124931	48617	3614	1922	11964
	LNP fit	632557	380696	272198	20610	13580	42883
Football	Linear fit	374850	202814	189570	37255	31949	48007
	Quadratic fit	84490	44399	45216	25268	22055	32642
	Exponential fit	463217	278921	203649	31563	27075	40760
	Logarithmic fit	437905	237276	220239	7754	7309	10395
	Power Regression	238426	151673	94171	22117	19846	21781
	LNP fit	798407	434374	394325	58754	49410	77004

Table B.20 – D-QP and R-D Mean Absolute Error (SSD_JND; IBBP GOP2, Akiyo, Foreman, Football)

Sequence	Fit Method	PSPNR-QP		Rate - PSPNR	
		I Type	P Type	I Type	P Type
Akiyo	Linear fit	0.26	0.15	1439	3199
	Quadratic fit	0.05	0.03	338	597
	Exponential fit	0.09	0.05	1065	4326
	Logarithmic fit	0.07	0.04	1847	5144
	Power Regression	0.14	0.08	681	1790
	LNP fit	1.43	0.85	5953	23453
Foreman	Linear fit	0.27	0.17	6225	8045
	Quadratic fit	0.06	0.04	902	1688
	Exponential fit	0.14	0.09	5650	20453
	Logarithmic fit	0.13	0.09	7743	11738
	Power Regression	0.10	0.06	3147	13462
	LNP fit	1.17	0.70	21121	53402
Football	Linear fit	0.34	0.22	8187	8110
	Quadratic fit	0.06	0.05	1644	3380
	Exponential fit	0.18	0.13	17363	28847
	Logarithmic fit	0.16	0.12	12247	13420
	Power Regression	0.08	0.05	10565	19096
	LNP fit	1.24	0.73	50256	71242

Table B.21 – D-QP and R-D Mean Absolute Error (PSPNR; IPPP GOP1, Akiyo, Foreman, Football)

Sequence	Fit Method	PSPNR-QP		Rate - PSPNR	
		I Type	P Type	I Type	P Type
Akiyo	Linear fit	0.44	0.14	1571	2115
	Quadratic fit	0.08	0.03	373	420
	Exponential fit	0.17	0.05	1637	2513
	Logarithmic fit	0.14	0.04	2189	3207
	Power Regression	0.19	0.07	800	1088
	LNP fit	2.24	0.76	8412	13756
Foreman	Linear fit	0.45	0.15	6340	6846
	Quadratic fit	0.10	0.03	899	1182
	Exponential fit	0.25	0.08	6604	12295
	Logarithmic fit	0.22	0.07	8069	9172
	Power Regression	0.15	0.05	3460	7943
	LNP fit	1.83	0.62	23488	34676
Football	Linear fit	0.53	0.19	8281	8162
	Quadratic fit	0.10	0.04	1886	2210
	Exponential fit	0.28	0.10	18699	20892
	Logarithmic fit	0.26	0.10	12511	12517
	Power Regression	0.13	0.04	11404	13316
	LNP fit	1.96	0.65	53407	56279

Table B.22 – D-QP and R-D Mean Absolute Error (PSPNR; IPPP GOP2, Akiyo, Foreman, Football)

Sequence	Fit Method	PSPNR-QP			Rate - PSPNR		
		I Type	P Type	B Type	I Type	P Type	B Type
Akiyo	Linear fit	0.79	0.26	0.18	1824	882	2031
	Quadratic fit	0.13	0.05	0.04	421	274	409
	Exponential fit	0.30	0.10	0.07	2406	342	1713
	Logarithmic fit	0.25	0.08	0.06	2888	1018	2781
	Power Regression	0.33	0.11	0.07	1101	432	757
	LNP fit	3.95	1.32	0.87	13119	2143	10194
Foreman	Linear fit	0.82	0.28	0.16	5714	5273	7326
	Quadratic fit	0.18	0.06	0.04	1336	1394	1834
	Exponential fit	0.46	0.15	0.08	4923	2086	9550
	Logarithmic fit	0.41	0.14	0.07	7315	6101	9354
	Power Regression	0.22	0.07	0.07	3224	2215	6220
	LNP fit	3.13	1.07	0.72	23455	12995	29515
Football	Linear fit	1.09	0.35	0.21	8018	7575	10206
	Quadratic fit	0.27	0.08	0.04	3294	1660	2444
	Exponential fit	0.66	0.21	0.11	27211	19392	20233
	Logarithmic fit	0.62	0.19	0.10	12239	11101	15389
	Power Regression	0.26	0.08	0.07	18060	12711	11815
	LNP fit	3.28	1.14	0.79	61337	43434	61898

Table B.23 – D-QP and R-D Mean Absolute Error (PSPNR; IBBP GOP1, Akiyo, Foreman, Football)

Sequence	Fit Method	PSNR-QP			Rate - PSNR		
		I Type	P Type	B Type	I Type	P Type	B Type
Akiyo	Linear fit	0.47	0.28	0.17	1390	839	2604
	Quadratic fit	0.10	0.06	0.04	349	255	523
	Exponential fit	0.17	0.11	0.06	1540	307	2504
	Logarithmic fit	0.14	0.09	0.05	2037	969	3754
	Power Regression	0.23	0.13	0.08	755	407	1045
	LNP fit	2.48	1.47	0.88	8530	2042	15027
Foreman	Linear fit	0.50	0.32	0.15	5567	5049	8155
	Quadratic fit	0.10	0.07	0.04	1228	1311	2329
	Exponential fit	0.27	0.18	0.07	3615	1922	11964
	Logarithmic fit	0.24	0.16	0.07	6755	5854	10757
	Power Regression	0.13	0.08	0.07	2337	2251	7846
	LNP fit	1.99	1.19	0.74	18384	12486	37135
Football	Linear fit	0.65	0.42	0.21	7754	7309	10395
	Quadratic fit	0.14	0.10	0.04	2344	1683	2678
	Exponential fit	0.37	0.26	0.11	22117	19847	21781
	Logarithmic fit	0.35	0.24	0.10	11622	10808	15914
	Power Regression	0.14	0.10	0.07	14412	13147	12855
	LNP fit	2.15	1.27	0.79	51068	43356	66449

Table B.24 – D-QP and R-D Mean Absolute Error (PSPNR; IBBP GOP2, Akiyo, Foreman, Football)

B.4 SSIM-QP curve fitting Tables (SSIM)

Sequence	Fit Method	IBBP GOP1			IBBP GOP2		
		I Type	P Type	B Type	I Type	P Type	B Type
AKI	Linear fit	345.4	3082.9	6980.4	887.3	2657.1	7296.9
	Quadratic fit	3.2	30.6	77.5	7.3	18.9	60.1
	Exponential fit	996.5	8840.3	20197.7	2511.6	7469.8	20646.2
	Logarithmic fit	804.4	7155.5	16178.9	2038.8	6078.2	16618.3
	Power Regression	1709.8	15152.7	34525.4	4285.0	12718.2	35003.7
	LNP fit	8452.5	74907.1	169153.7	20994.6	62203.0	169038.4
CGD	Linear fit	533.1	4161.5	16612.8	1237.3	2748.1	15098.6
	Quadratic fit	258.6	1914.4	2725.8	554.1	1422.1	2442.9
	Exponential fit	8438.5	63526.1	93131.4	18021.8	50739.9	85232.6
	Logarithmic fit	335.3	2688.6	6152.9	726.0	2276.3	5501.3
	Power Regression	14344.2	109773.8	180753.7	31160.0	87034.3	166600.1
	LNP fit	17882.5	153404.8	276188.5	42524.0	131080.1	274409.4
DEA	Linear fit	376.8	3546.7	7215.8	1007.3	3177.6	7699.9
	Quadratic fit	13.8	105.7	316.0	26.7	71.1	293.4
	Exponential fit	2632.5	23324.1	53554.0	6578.6	19864.1	54397.0
	Logarithmic fit	1079.2	9892.6	21068.2	2799.7	8623.9	21893.3
	Power Regression	4302.3	37957.7	87458.5	10689.6	32110.9	88335.8
	LNP fit	15058.6	133234.0	301101.4	37427.8	111883.9	302460.9
FLG	Linear fit	411.0	3442.1	5756.6	926.9	3225.2	6324.5
	Quadratic fit	67.0	588.0	1347.5	144.6	386.7	1099.4
	Exponential fit	4295.2	36381.9	82340.9	9370.6	27981.7	74321.3
	Logarithmic fit	1183.7	10129.2	18942.3	2768.9	9053.2	20046.2
	Power Regression	6857.5	58426.8	133869.2	15134.5	44775.6	120606.9
	LNP fit	18456.8	161217.8	356163.5	43869.9	131034.4	343746.5
FOR	Linear fit	127.5	1075.4	2572.1	236.1	654.4	2042.5
	Quadratic fit	35.0	429.1	1112.1	98.7	237.9	937.9
	Exponential fit	1475.1	16666.7	45307.5	4019.4	13377.0	44943.7
	Logarithmic fit	178.7	2053.8	5291.9	555.8	1727.4	5813.6
	Power Regression	2919.3	31359.8	82533.6	7672.3	24943.0	80533.4
	LNP fit	9697.0	94901.4	225613.3	25809.5	79582.5	230203.6
FOT	Linear fit	1046.2	8018.2	22824.1	2106.0	5821.5	20530.8
	Quadratic fit	123.9	1192.1	3953.5	330.3	1119.6	4290.6
	Exponential fit	2276.6	24609.8	69436.7	6802.5	22483.7	75421.0
	Logarithmic fit	479.7	3682.9	10950.7	995.0	2677.6	9800.9
	Power Regression	5019.7	51254.8	141997.4	14007.5	45397.0	149778.0
	LNP fit	10579.5	99716.4	220447.3	27911.8	89012.1	230349.6

Table B.25 – Cumulative squared error for SSIM-QP curve fitting for video sequences Akiyo, Coastguard, Deadline, Flower Garden, Foreman and Football (IBBP GOP1 and IBBP GOP2)

Sequence	Fit Method	IBBP GOP1			IBBP GOP2		
		I Type	P Type	B Type	I Type	P Type	B Type
HAL	Linear fit	221.6	1860.7	5186.6	515.1	1464.3	5042.9
	Quadratic fit	9.7	83.4	226.0	24.7	79.4	233.4
	Exponential fit	603.2	5076.3	12609.8	1394.7	4123.4	12253.1
	Logarithmic fit	507.5	4313.9	10920.8	1192.7	3481.1	10656.6
	Power Regression	1034.3	8773.8	20852.8	2412.3	7208.8	20315.0
	LNP fit	5404.6	46827.0	102213.2	12905.2	39407.8	100215.0
MAD	Linear fit	277.0	2552.0	5520.5	744.2	2290.2	6031.5
	Quadratic fit	40.4	322.1	1143.0	100.9	266.5	1190.0
	Exponential fit	1935.8	16967.1	38720.2	4812.1	14193.9	39209.8
	Logarithmic fit	833.9	7544.8	16371.7	2158.5	6553.4	17205.8
	Power Regression	3295.3	28826.1	66106.1	8151.7	23994.8	66477.9
	LNP fit	12804.0	112583.5	254408.5	31671.5	93837.5	254498.9
NEW	Linear fit	403.8	3501.1	7735.0	982.0	2997.9	7866.9
	Quadratic fit	21.9	142.9	438.1	37.8	114.5	411.5
	Exponential fit	2136.4	17741.5	42778.0	5026.9	14790.4	42759.3
	Logarithmic fit	1064.6	9199.3	20678.8	2587.1	7790.8	20874.2
	Power Regression	3522.3	29376.8	70516.9	8317.2	24454.5	70394.7
	LNP fit	13548.1	115968.3	268371.1	32685.8	96641.4	268021.6
PAR	Linear fit	119.7	1157.6	2572.8	323.9	1025.4	2698.5
	Quadratic fit	27.9	242.1	633.4	66.8	200.3	615.1
	Exponential fit	2176.4	19314.3	48280.8	5392.9	15705.6	47492.4
	Logarithmic fit	600.7	5580.4	12593.8	1563.6	4778.8	12884.2
	Power Regression	3886.7	34382.0	84873.0	9605.2	28008.0	83525.6
	LNP fit	14705.3	130874.5	302895.6	36626.4	108624.6	302058.0
SIL	Linear fit	117.4	1001.6	2884.8	256.9	739.4	2479.7
	Quadratic fit	63.1	549.2	1561.1	141.3	420.4	1498.5
	Exponential fit	2481.3	22201.2	53224.1	6121.5	18691.4	54478.2
	Logarithmic fit	156.8	1377.4	3227.1	389.4	1217.4	3543.1
	Power Regression	4770.9	42641.9	101705.4	11764.8	35755.3	103306.8
	LNP fit	13877.1	124050.2	280543.8	34706.9	104939.7	287722.8
MCL	Linear fit	369.7	3571.2	3857.4	872.6	3154.7	4674.7
	Quadratic fit	62.0	402.4	954.3	116.9	311.5	780.7
	Exponential fit	4772.1	41875.3	86596.7	9927.6	33308.2	79423.1
	Logarithmic fit	1168.3	10909.3	16128.6	2762.0	9327.9	17514.4
	Power Regression	7503.4	65500.4	140863.3	15891.8	52224.8	129198.4
	LNP fit	19220.0	170326.1	364710.6	45230.5	140903.9	356522.4

Table B.26 – Cumulative squared error for SSIM-QP curve fitting for video sequences Hall, Mother and Daughter, News, Paris, Silence and Mobile and Calendar (IBBP GOP1 and IBBP GOP2)

Sequence	Fit Method	IPPP GOP1		IPPP GOP2	
		I Type	P Type	I Type	P Type
AKI	Linear fit	2482.5	7429.5	967.7	8877.6
	Quadratic fit	33.2	102.2	13.2	106.8
	Exponential fit	7421.9	22049.9	2967.8	26609.7
	Logarithmic fit	5891.3	17564.7	2328.3	21068.5
	Power Regression	12805.7	38010.3	5141.0	45877.6
	LNP fit	63775.1	189115.3	25728.8	228078.4
CGD	Linear fit	3966.0	6955.1	1949.7	13582.7
	Quadratic fit	1565.4	3431.2	776.0	5535.1
	Exponential fit	48395.9	136519.2	23404.7	181249.2
	Logarithmic fit	2067.2	5816.1	1012.1	7872.5
	Power Regression	85719.0	236066.9	40797.5	317491.8
	LNP fit	121921.9	377931.8	50200.5	444678.6
DEA	Linear fit	2915.4	9203.7	1093.3	10255.6
	Quadratic fit	94.1	267.7	45.2	341.9
	Exponential fit	20371.3	61107.9	8190.1	72864.7
	Logarithmic fit	8280.1	25521.1	3204.3	29346.3
	Power Regression	33091.4	98894.7	13363.6	118475.3
	LNP fit	113912.2	340536.4	45839.8	407346.5
FLG	Linear fit	2677.3	9803.2	1071.2	9940.2
	Quadratic fit	445.7	1203.1	242.2	1806.4
	Exponential fit	30129.5	91326.3	14056.2	113207.3
	Logarithmic fit	8195.1	27704.0	3296.7	29985.8
	Power Regression	48519.1	145073.7	22466.8	181530.6
	LNP fit	134569.4	404916.7	56963.8	490195.3
FOR	Linear fit	749.9	1994.4	390.2	2526.0
	Quadratic fit	256.0	714.6	113.8	986.7
	Exponential fit	14056.9	43182.2	5509.5	52222.5
	Logarithmic fit	1720.8	5289.0	645.4	6033.7
	Power Regression	26203.9	79521.3	10499.1	96746.5
	LNP fit	80371.7	241975.3	31656.4	290285.8
FOT	Linear fit	5793.9	15022.3	2618.4	21542.7
	Quadratic fit	1111.7	3118.7	441.7	3799.1
	Exponential fit	23072.6	66364.8	8917.5	78232.9
	Logarithmic fit	2770.9	6848.6	1282.0	10187.2
	Power Regression	46140.0	133456.0	18273.9	158740.9
	LNP fit	86415.2	264255.5	33649.8	303623.7

Table B.27 – Cumulative squared error for SSIM-QP curve fitting for video sequences Akiyo, Coastguard, Deadline, Flower Garden, Foreman and Football (IPPP GOP1 and IPPP GOP2)

Sequence	Fit Method	IPPP GOP1		IPPP GOP2	
		I Type	P Type	I Type	P Type
HAL	Linear fit	1543.4	4350.7	634.0	5443.0
	Quadratic fit	75.1	249.8	26.1	264.2
	Exponential fit	4263.1	12548.9	1768.3	15319.4
	Logarithmic fit	3587.4	10423.1	1474.7	12792.3
	Power Regression	7359.8	21940.9	3047.4	26553.1
	LNP fit	38989.1	119279.0	15990.9	141467.1
MAD	Linear fit	2015.0	5974.1	723.3	6890.8
	Quadratic fit	317.5	914.1	121.0	1056.6
	Exponential fit	14108.1	40791.4	5924.9	50716.1
	Logarithmic fit	6096.1	18069.4	2331.3	21356.9
	Power Regression	24121.1	70054.4	10150.2	86889.8
	LNP fit	94396.5	279113.4	38748.1	338901.8
NEW	Linear fit	2973.8	8975.0	1109.2	10511.3
	Quadratic fit	125.8	330.7	55.3	460.4
	Exponential fit	15971.0	46660.8	6432.2	56963.1
	Logarithmic fit	7897.3	23613.6	3048.3	28023.8
	Power Regression	26319.8	76899.2	10655.4	93924.1
	LNP fit	100828.2	297034.3	40735.2	359896.7
PAR	Linear fit	972.6	2971.5	375.2	3466.5
	Quadratic fit	229.7	618.1	95.8	797.2
	Exponential fit	16738.2	48512.7	7089.9	60314.4
	Logarithmic fit	4697.7	14178.6	1858.2	16793.8
	Power Regression	29730.4	86339.7	12515.2	106915.9
	LNP fit	111290.4	329499.7	45227.1	398045.9
SIL	Linear fit	1262.1	3487.1	576.1	4491.3
	Quadratic fit	462.4	1275.0	201.6	1626.3
	Exponential fit	18346.7	53411.2	7528.8	65402.0
	Logarithmic fit	833.7	2445.5	334.8	2949.9
	Power Regression	36013.8	105156.3	14779.2	128406.2
	LNP fit	100309.3	300613.2	40013.8	357974.9
MCL	Linear fit	2607.8	9062.1	1001.5	9489.9
	Quadratic fit	464.3	1401.0	229.4	1841.6
	Exponential fit	42113.1	126301.0	18635.7	153566.4
	Logarithmic fit	8649.3	28056.7	3413.3	31168.8
	Power Regression	65528.9	195796.5	28805.2	238740.9
	LNP fit	151569.1	459234.3	62122.0	545774.3

Table B.28 – Cumulative squared error for SSIM-QP curve fitting for video sequences Hall, Mother and Daughter, News, Paris, Silence and Mobile and Calendar (IPPP GOP1 and IPPP GOP2)

B.5 Rate-SSIM Curve Fitting Tables (SSIM)

Sequence	Fit Method	IBBP GOP1			IBBP GOP2		
		I Type	P Type	B Type	I Type	P Type	B Type
AKI	Linear fit	3,01E+08	1,61E+08	4,31E+09	3,38E+08	1,21E+08	8,53E+09
	Quadratic fit	4,74E+07	5,30E+07	7,85E+08	5,95E+07	3,99E+07	1,39E+09
	Exponential fit	8,66E+07	8,93E+07	1,38E+09	1,05E+08	6,74E+07	2,57E+09
	Logarithmic fit	4,47E+08	1,94E+08	6,11E+09	4,91E+08	1,45E+08	1,25E+10
	Power Regression	1,96E+08	1,34E+08	2,91E+09	2,24E+08	1,01E+08	5,73E+09
	LNP fit	2,33E+09	4,63E+08	2,72E+10	2,44E+09	3,46E+08	6,03E+10
CGD	Linear fit	4,18E+09	3,41E+10	6,57E+10	9,52E+09	2,84E+10	6,77E+10
	Quadratic fit	4,20E+08	3,67E+09	3,98E+09	8,65E+08	3,03E+09	3,91E+09
	Exponential fit	5,47E+08	3,00E+09	1,49E+11	1,71E+09	1,71E+09	1,49E+11
	Logarithmic fit	1,21E+10	7,18E+10	2,86E+11	2,37E+10	5,85E+10	3,01E+11
	Power Regression	6,37E+09	4,52E+10	6,19E+10	1,31E+10	3,78E+10	6,31E+10
	LNP fit	2,35E+10	1,23E+11	7,55E+11	4,47E+10	9,95E+10	8,37E+11
DEA	Linear fit	1,40E+09	2,17E+09	1,88E+10	1,80E+09	1,81E+09	3,42E+10
	Quadratic fit	1,80E+08	4,85E+08	2,62E+09	2,69E+08	4,25E+08	4,17E+09
	Exponential fit	2,18E+08	6,06E+08	2,83E+09	3,22E+08	5,48E+08	4,35E+09
	Logarithmic fit	2,71E+09	3,17E+09	3,65E+10	3,28E+09	2,60E+09	7,05E+10
	Power Regression	1,13E+09	1,84E+09	1,58E+10	1,46E+09	1,55E+09	2,88E+10
	LNP fit	9,93E+09	6,91E+09	1,24E+11	1,12E+10	5,61E+09	2,61E+11
FLG	Linear fit	1,10E+10	8,36E+10	2,09E+11	2,66E+10	6,68E+10	2,30E+11
	Quadratic fit	2,03E+09	1,78E+10	2,81E+10	5,09E+09	1,59E+10	3,13E+10
	Exponential fit	2,47E+09	2,49E+10	2,81E+10	6,51E+09	2,46E+10	3,36E+10
	Logarithmic fit	1,85E+10	1,25E+11	4,23E+11	4,17E+10	9,50E+10	4,73E+11
	Power Regression	1,08E+10	8,68E+10	1,86E+11	2,63E+10	6,95E+10	2,04E+11
	LNP fit	4,90E+10	2,31E+11	1,12E+12	9,24E+10	1,75E+11	1,38E+12
FOR	Linear fit	7,64E+08	4,29E+09	2,29E+10	1,59E+09	3,37E+09	3,13E+10
	Quadratic fit	8,54E+07	8,55E+08	2,59E+09	2,37E+08	6,82E+08	3,15E+09
	Exponential fit	1,18E+08	1,20E+09	4,26E+09	3,09E+08	1,06E+09	5,45E+09
	Logarithmic fit	1,43E+09	6,73E+09	5,46E+10	2,71E+09	5,21E+09	8,20E+10
	Power Regression	5,77E+08	4,50E+09	2,38E+10	1,41E+09	3,55E+09	3,31E+10
	LNP fit	5,67E+09	1,51E+10	1,51E+11	8,67E+09	1,16E+10	2,38E+11
FOT	Linear fit	1,47E+09	1,02E+10	3,01E+10	3,39E+09	9,28E+09	3,42E+10
	Quadratic fit	1,79E+08	1,04E+09	6,41E+09	3,86E+08	9,75E+08	7,00E+09
	Exponential fit	1,63E+09	1,01E+10	3,82E+10	2,91E+09	6,77E+09	3,77E+10
	Logarithmic fit	7,87E+09	4,16E+10	1,70E+11	1,50E+10	3,53E+10	1,96E+11
	Power Regression	1,68E+09	9,05E+09	4,57E+10	3,45E+09	8,17E+09	5,52E+10
	LNP fit	2,87E+10	1,32E+11	5,81E+11	5,05E+10	1,09E+11	6,64E+11

Table B.29 – Cumulative squared error for Rate-SSIM curve fitting for video sequences Akiyo, Coastguard, Deadline, Flower Garden, Foreman and Football (IBBP GOP1 and IBBP GOP2)

Sequence	Fit Method	IBBP GOP1			IBBP GOP2		
		I Type	P Type	B Type	I Type	P Type	B Type
HAL	Linear fit	1,11E+09	7,42E+09	3,04E+10	2,34E+09	7,19E+09	3,99E+10
	Quadratic fit	2,53E+08	2,50E+09	7,26E+09	6,76E+08	2,41E+09	8,41E+09
	Exponential fit	4,23E+08	3,88E+09	1,27E+10	1,10E+09	3,68E+09	1,54E+10
	Logarithmic fit	1,46E+09	8,64E+09	3,83E+10	2,89E+09	8,38E+09	5,21E+10
	Power Regression	7,88E+08	6,09E+09	2,21E+10	1,82E+09	5,87E+09	2,82E+10
	LNP fit	5,42E+09	1,65E+10	9,71E+10	8,15E+09	1,62E+10	1,51E+11
MAD	Linear fit	6,37E+08	7,89E+08	1,13E+10	7,68E+08	6,50E+08	1,79E+10
	Quadratic fit	1,02E+08	2,03E+08	2,30E+09	1,38E+08	1,63E+08	3,45E+09
	Exponential fit	8,30E+07	2,03E+08	2,28E+09	1,25E+08	1,65E+08	3,30E+09
	Logarithmic fit	1,13E+09	1,11E+09	1,90E+10	1,30E+09	9,22E+08	3,14E+10
	Power Regression	4,17E+08	5,64E+08	8,50E+09	5,18E+08	4,64E+08	1,33E+10
	LNP fit	4,31E+09	2,58E+09	6,02E+10	4,69E+09	2,16E+09	1,11E+11
NEW	Linear fit	6,61E+08	8,75E+08	9,93E+09	7,81E+08	7,46E+08	1,71E+10
	Quadratic fit	9,77E+07	1,59E+08	1,55E+09	1,19E+08	1,43E+08	2,51E+09
	Exponential fit	1,27E+08	1,52E+08	1,94E+09	1,51E+08	1,29E+08	3,34E+09
	Logarithmic fit	1,19E+09	1,36E+09	1,77E+10	1,37E+09	1,16E+09	3,16E+10
	Power Regression	5,10E+08	5,48E+08	7,84E+09	5,89E+08	4,59E+08	1,39E+10
	LNP fit	4,64E+09	4,01E+09	6,27E+10	5,13E+09	3,48E+09	1,19E+11
PAR	Linear fit	1,69E+09	4,78E+09	2,63E+10	2,51E+09	3,92E+09	4,14E+10
	Quadratic fit	1,54E+08	5,65E+08	2,41E+09	2,43E+08	4,65E+08	3,62E+09
	Exponential fit	4,25E+07	1,48E+08	6,53E+08	7,48E+07	1,21E+08	1,04E+09
	Logarithmic fit	3,97E+09	8,52E+09	6,34E+10	5,44E+09	6,91E+09	1,07E+11
	Power Regression	1,21E+09	2,41E+09	1,91E+10	1,62E+09	2,00E+09	3,29E+10
	LNP fit	1,77E+10	2,45E+10	2,49E+11	2,20E+10	2,00E+10	4,74E+11
SIL	Linear fit	7,71E+08	7,02E+08	9,90E+09	8,41E+08	6,35E+08	2,08E+10
	Quadratic fit	3,72E+07	7,74E+07	7,62E+08	4,53E+07	6,91E+07	1,45E+09
	Exponential fit	1,77E+08	3,48E+07	7,53E+08	1,80E+08	2,86E+07	1,74E+09
	Logarithmic fit	2,90E+09	1,62E+09	3,23E+10	3,05E+09	1,46E+09	7,02E+10
	Power Regression	7,17E+08	5,80E+08	9,66E+09	7,83E+08	5,17E+08	2,02E+10
	LNP fit	8,89E+09	3,77E+09	9,41E+10	9,20E+09	3,40E+09	2,09E+11
MCL	Linear fit	1,94E+10	1,62E+11	3,01E+11	4,57E+10	1,30E+11	3,21E+11
	Quadratic fit	3,53E+09	3,06E+10	3,48E+10	8,46E+09	2,58E+10	3,77E+10
	Exponential fit	2,76E+09	2,85E+10	2,15E+10	7,22E+09	2,46E+10	2,77E+10
	Logarithmic fit	3,41E+10	2,50E+11	6,15E+11	7,36E+10	1,98E+11	6,83E+11
	Power Regression	1,97E+10	1,65E+11	2,89E+11	4,37E+10	1,28E+11	3,12E+11
	LNP fit	7,87E+10	4,37E+11	1,27E+12	1,59E+11	3,69E+11	1,55E+12

Table B.30 – Cumulative squared error for Rate-SSIM curve fitting for video sequences Hall, Mother and Daughter, News, Paris, Silence and Mobile and Calendar (IBBP GOP1 and IBBP GOP2)

Sequence	Fit Method	IPPP GOP1		IPPP GOP2	
		I Type	P Type	I Type	P Type
AKI	Linear fit	613473388	20707794920	415092949	9002022307
	Quadratic fit	127297508	3047888918	73668532	1401881644
	Exponential fit	232439444	5606145433	127498323	2581468800
	Logarithmic fit	840719008	31082809444	595367424	13304165730
	Power Regression	439208757	13431156922	274385961	5930714565
	LNP fit	3438165192	165379107560	2811032072	67964689040
CGD	Linear fit	18229867587	92314822430	7065143729	78208975418
	Quadratic fit	1231841802	6427244511	529878546	5465990489
	Exponential fit	10054563686	27659706746	3856472759	33316574702
	Logarithmic fit	65343482859	372800519067	27926530810	297942079209
	Power Regression	18176966131	123435780295	8554435593	92653978118
	LNP fit	150702256491	950242027780	64491561006	719042057104
DEA	Linear fit	3372260310	85736918172	2005688172	39058837635
	Quadratic fit	516666876	9286972100	272264569	4531465213
	Exponential fit	661567070	9909579785	324502473	5131165055
	Logarithmic fit	5905579111	179480922162	3753987736	79013926035
	Power Regression	2792686918	70431795513	1641322618	32207873613
	LNP fit	18021360396	699103587039	12695503614	294534435908
FLG	Linear fit	65014197385	323045229512	26564469176	278150019353
	Quadratic fit	8366853301	41428223153	3420755488	35041712063
	Exponential fit	4290171237	30439360333	1575639394	20243607189
	Logarithmic fit	120462553753	736981040279	52492205022	577447491453
	Power Regression	45981236341	286898505754	20569033259	218808153325
	LNP fit	303889417534	2467034298954	137185734463	1667539035528
FOR	Linear fit	6039087836	53107969049	2619151790	34687045565
	Quadratic fit	594496123	3862827094	223832004	2977495454
	Exponential fit	699101874	12007842129	200107878	6161290425
	Logarithmic fit	12153398136	166806545398	5617666663	93209387258
	Power Regression	6363355938	54877367941	2612596186	36886609280
	LNP fit	32039016674	553736835541	15564683509	285732671902
FOT	Linear fit	9818632608	43256349336	4187542067	38943522169
	Quadratic fit	1225726656	5240157944	587513641	4915863749
	Exponential fit	7509860991	37978239014	2959054613	31895893340
	Logarithmic fit	45124691509	244275006152	19684322086	192430485244
	Power Regression	12453524156	63517725387	5395127371	51633874368
	LNP fit	146882389706	858076978653	66242431323	650103497582

Table B.31 – Cumulative squared error for Rate-SSIM curve fitting for video sequences Akiyo, Coastguard, Deadline, Flower Garden, Foreman and Football (IPPP GOP1 and IPPP GOP2)

Sequence	Fit Method	IPPP GOP1		IPPP GOP2	
		I Type	P Type	I Type	P Type
HAL	Linear fit	9538834718	60062458899	3619366025	44377452434
	Quadratic fit	2279018795	8343308829	855273376	8503823034
	Exponential fit	3952643739	16694857324	1433603149	15537497711
	Logarithmic fit	11789367661	86371196539	4567662753	59349905986
	Power Regression	7034236812	39076682731	2609714691	30992782366
	LNP fit	27974235748	368978311459	12257903520	196356316534
MAD	Linear fit	2061700134	36333245069	1079290656	18330104394
	Quadratic fit	447896611	5543919826	192912607	3109392286
	Exponential fit	562003282	4815356505	203373345	3098214396
	Logarithmic fit	3220696578	67514857289	1815444327	32691637101
	Power Regression	1612611822	25315911058	789873214	13311974449
	LNP fit	9564234579	266889819323	6131022588	120136728184
NEW	Linear fit	1718905536	40296939747	1128259443	18557971400
	Quadratic fit	298434509	5235608981	174763713	2562358615
	Exponential fit	417483183	7753723054	243045060	3759675887
	Logarithmic fit	2862210497	74802039818	1972236831	33534976387
	Power Regression	1368074055	32219734376	907153363	14736475485
	LNP fit	9231061879	299083797426	6754811461	129118893299
PAR	Linear fit	5217904403	90038407371	2719140584	45353483319
	Quadratic fit	474932244	7240745522	241922684	3766809810
	Exponential fit	144340206	2285333102	76738097	1133238864
	Logarithmic fit	11213854617	240824847958	6287366136	115160793252
	Power Regression	3505371990	72573200296	1984562608	34944138274
	LNP fit	42278772255	1192209900695	25293761140	537136875068
SIL	Linear fit	1443506402	48776623911	946977088	21281338990
	Quadratic fit	101614673	2260051126	54586640	1098821935
	Exponential fit	235614355	7892045543	209547777	3218394492
	Logarithmic fit	4894828182	183787587939	3484742464	78470429024
	Power Regression	1392402446	48171446455	918393768	21116988451
	LNP fit	14266160840	568735158433	10558100979	239485078073
MCL	Linear fit	106586315230	456365198116	41944450709	428641851422
	Quadratic fit	11988999358	50461120174	4384832556	46300223060
	Exponential fit	4360661602	33583455443	1079902494	20887773662
	Logarithmic fit	215953724052	1175260310112	91155030869	974493290466
	Power Regression	93216543824	462685399217	38041123795	395885313929
	LNP fit	420629953492	2809815466379	182681118300	2099757222355

Table B.32 – Cumulative squared error for Rate-SSIM curve fitting for video sequences Hall, Mother and Daughter, News, Paris, Silence and Mobile and Calendar (IPPP GOP1 and IPPP GOP2)

Annex C. Joint Coding Results

C.1 Joint Coding Results (Charts)

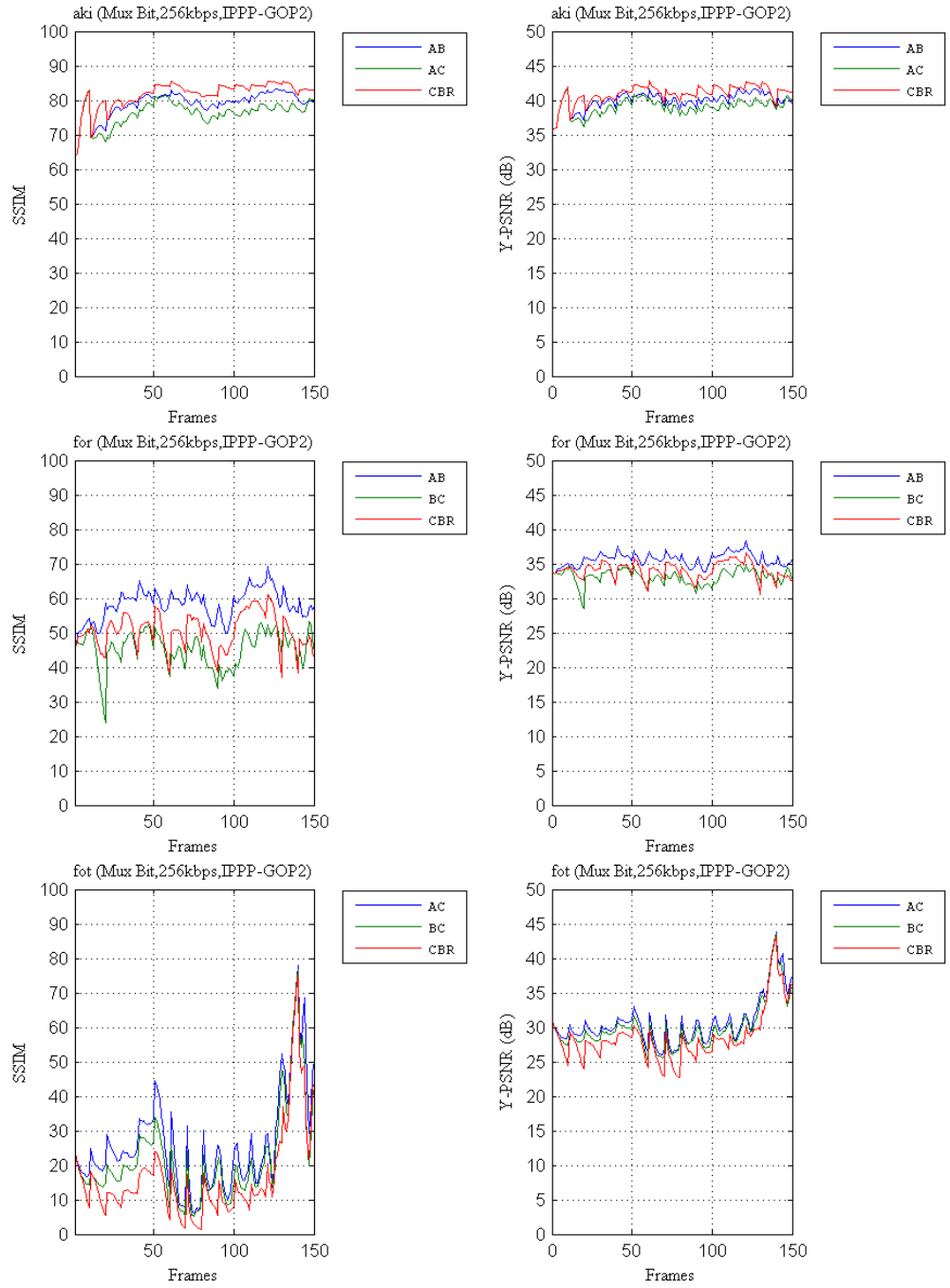


Figure C.1 – Joint Coding Mux Bit (IPPP GOP2; 256kbps; 2SRC)

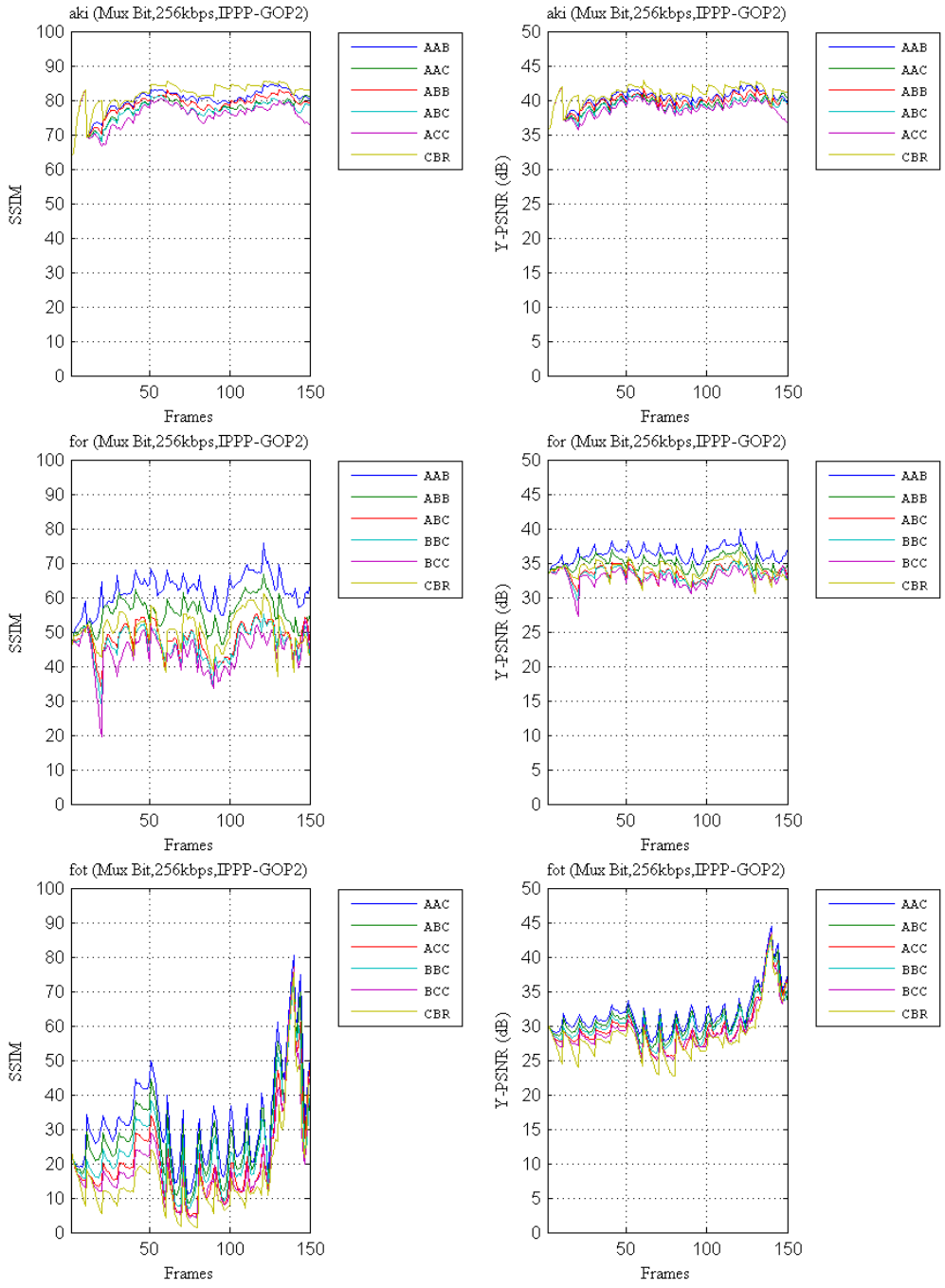


Figure C.2 – Joint Coding Mux Bit (IPPP GOP2; 256kbps; 3SRC)

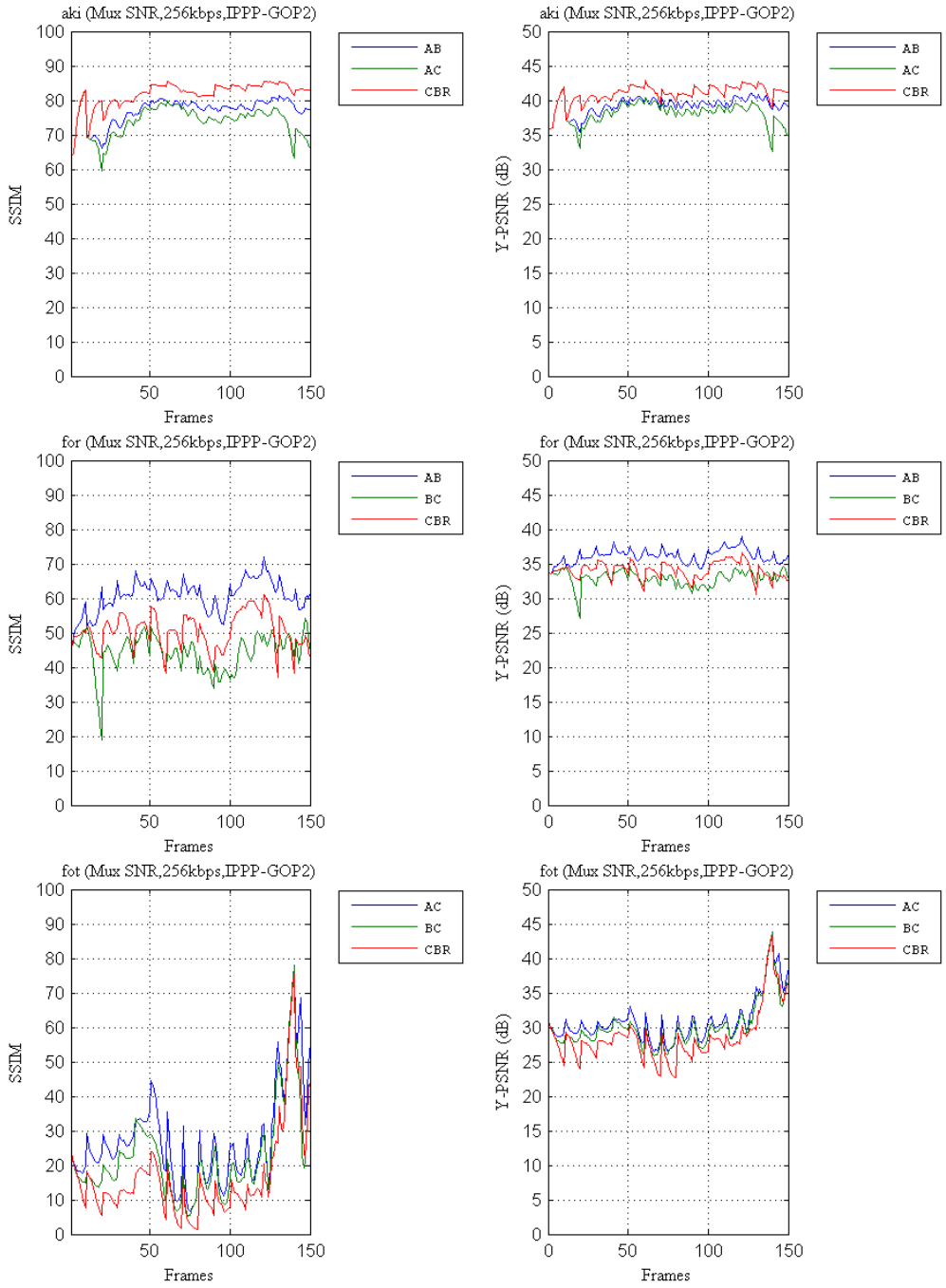


Figure C.3 – Joint Coding Mux PSNR (IPPP GOP2; 256kbps; 2SRC)

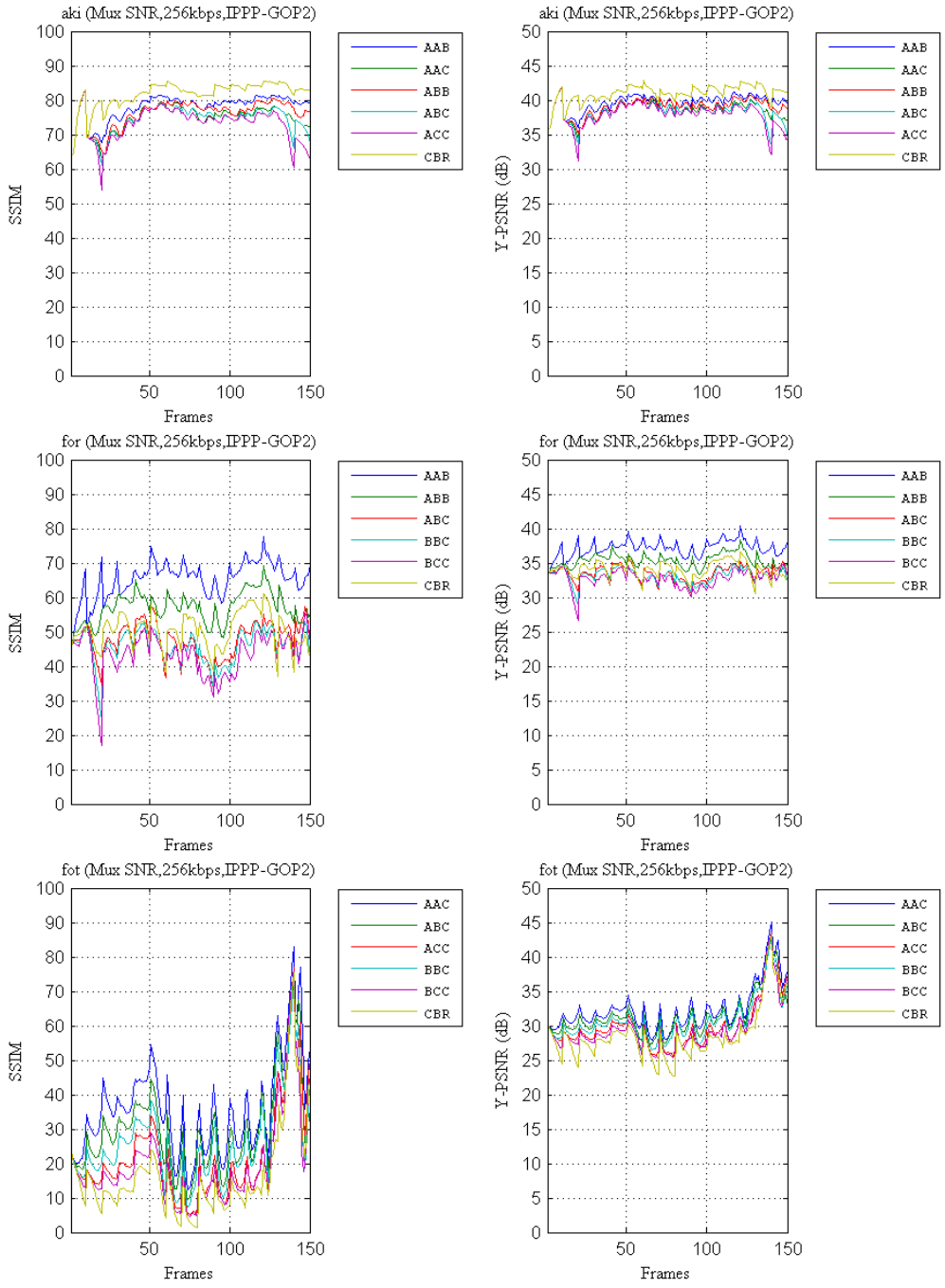


Figure C.4 – Joint Coding Mux PSNR (IPPP GOP2; 256kbps; 3SRC)

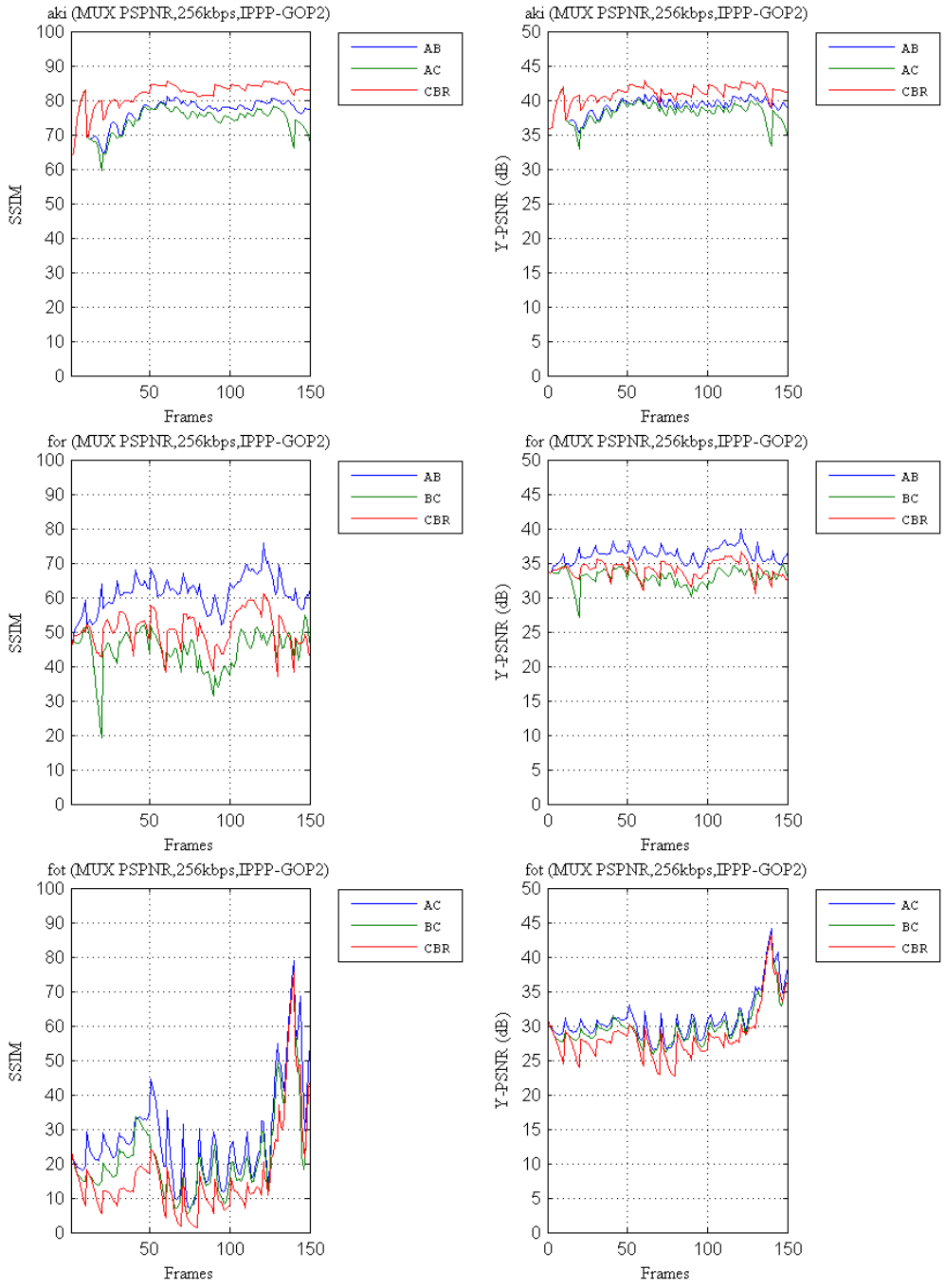


Figure C.5 – Joint Coding Mux PSPNR (IPPP GOP2; 256kbps; 2SRC)

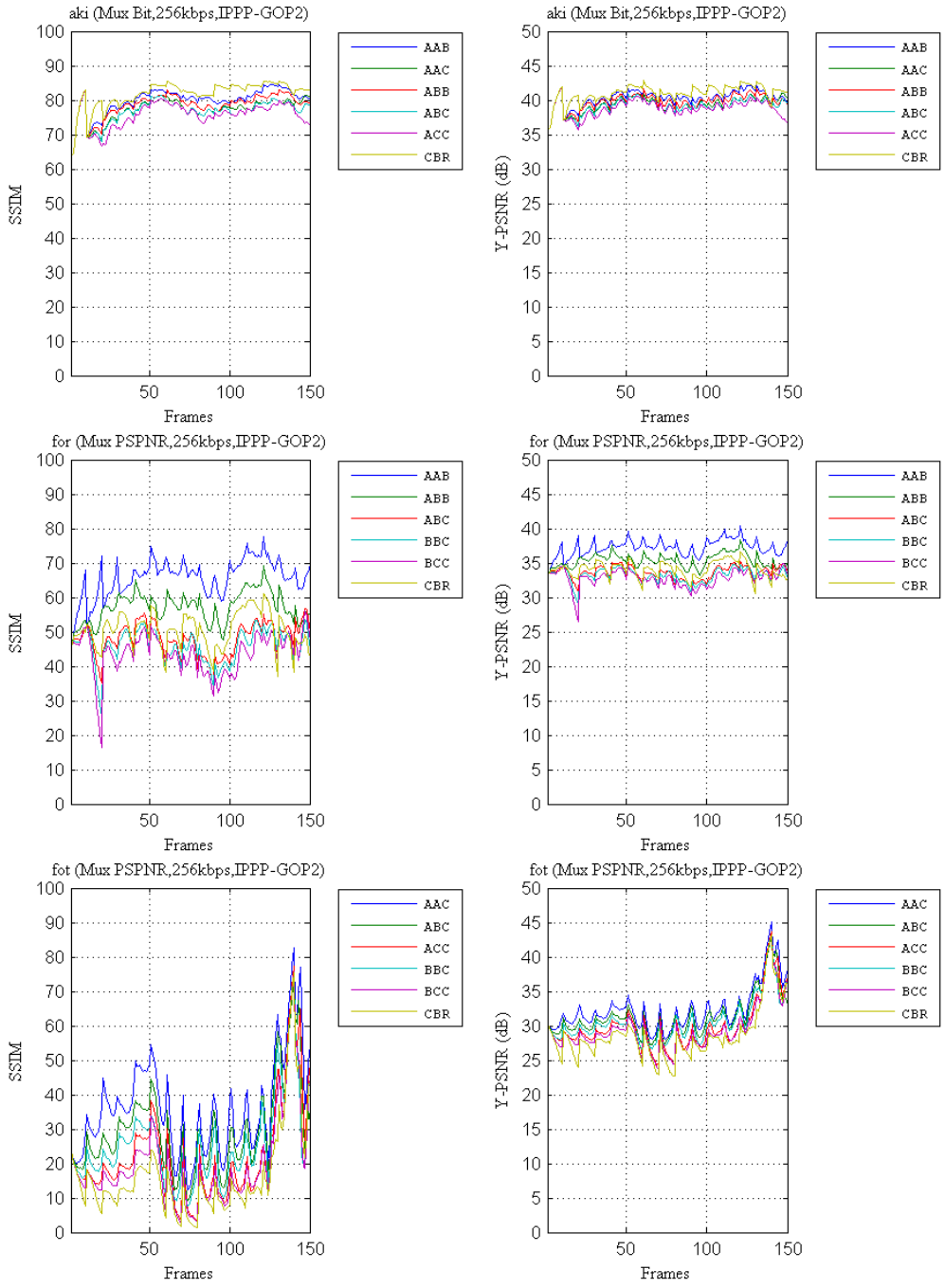


Figure C.6 – Joint Coding Mux PSPNR (IPPP GOP2; 256kbps; 3SRC)

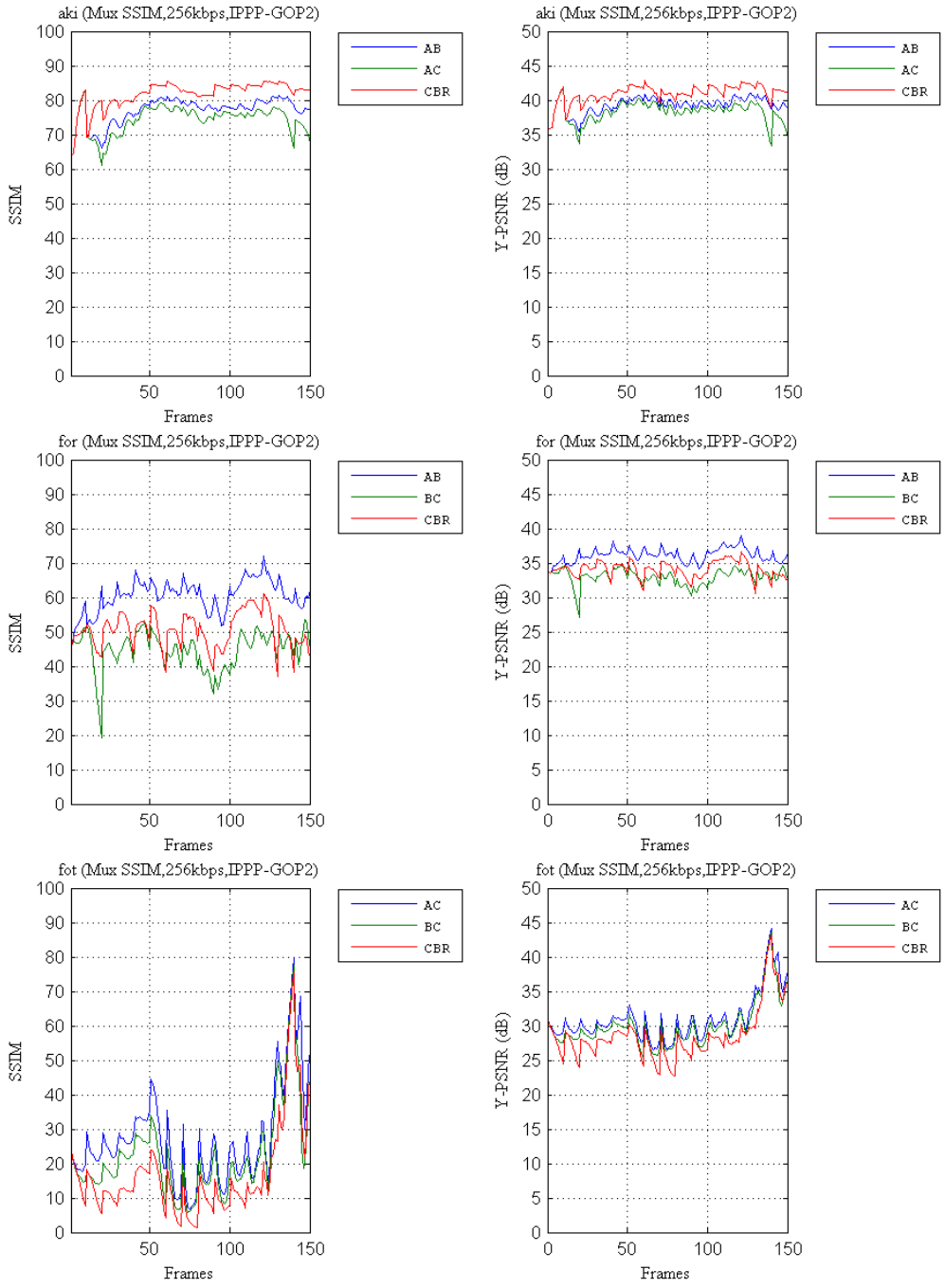


Figure C.7 – Joint Coding Mux SSIM (IPPP GOP2; 256kbps; 2SRC)

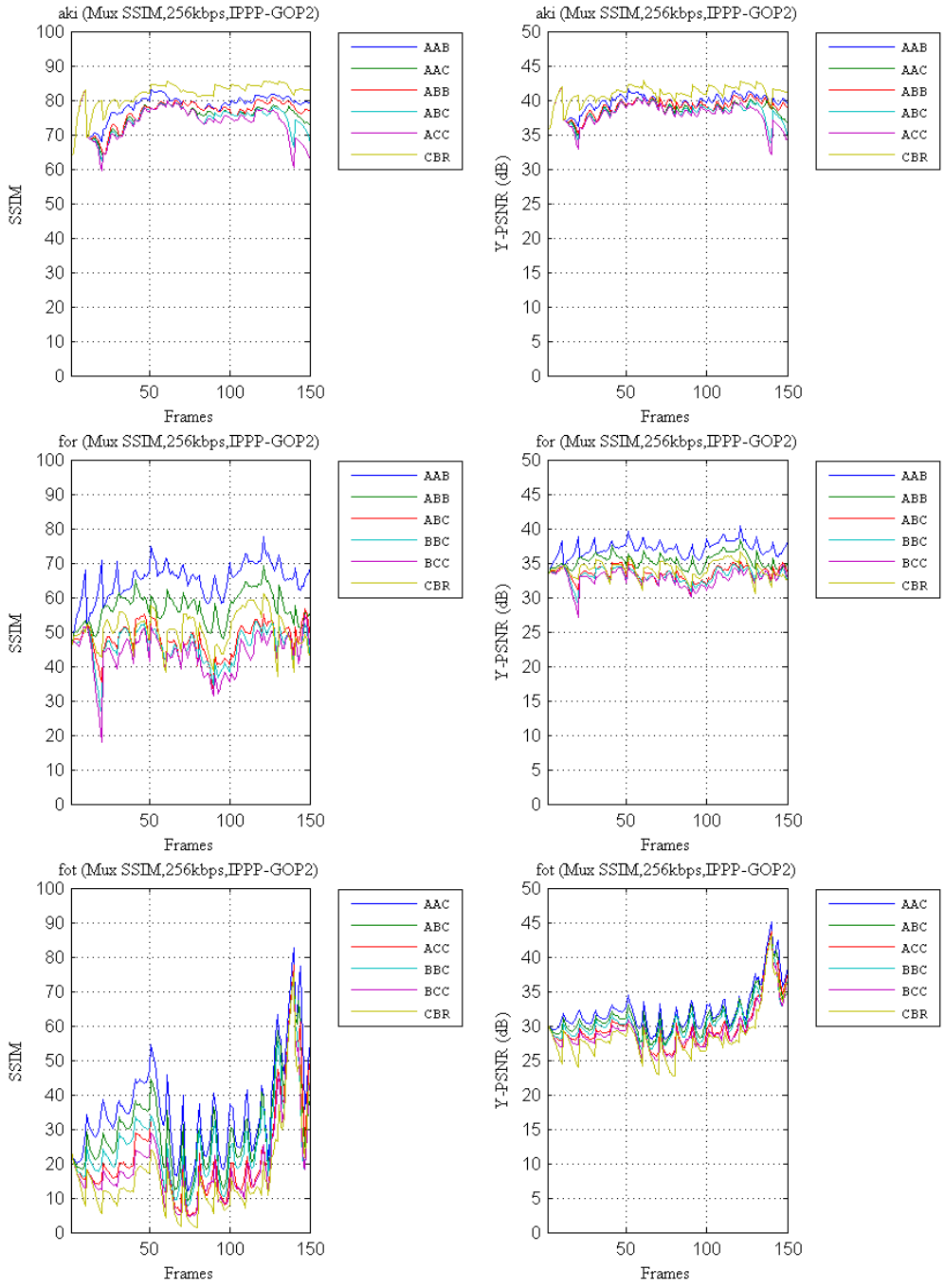


Figure C.8 – Joint Coding Mux SSIM (IPPP GOP2; 256kbps; 3SRC)

C.2 Joint Coding Results (Tables)

Bit Rate	Group Name	Metric	mux_bit	mux_psnr	mux_ssim	mux_pspnr
256	AB	PSNR	0.09	0.08	0.07	-0.05
256	AB	PSPNR	-0.20	-0.25	-0.27	-1.30
256	AB	SSIM	1.49	1.43	1.38	0.15
256	AC	PSNR	-0.79	-0.83	-0.87	-1.88
256	AC	PSPNR	-1.13	-1.16	-1.22	-2.71
256	AC	SSIM	-0.28	-0.30	-0.30	-0.57
256	BC	PSNR	0.13	0.11	0.15	-0.01
256	BC	PSPNR	0.11	0.10	0.14	-0.02
256	BC	SSIM	0.87	0.74	0.92	0.34
512	AB	PSNR	0.15	0.12	0.12	-0.03
512	AB	PSPNR	-0.24	-0.35	-0.39	-1.35
512	AB	SSIM	2.69	2.65	2.71	0.72
512	AC	PSNR	-0.13	-0.18	-0.23	-0.70
512	AC	PSPNR	-1.19	-1.35	-1.49	-2.88
512	AC	SSIM	3.64	3.76	3.57	2.60
512	BC	PSNR	0.12	0.11	0.12	-0.02
512	BC	PSPNR	0.09	0.08	0.10	-0.04
512	BC	SSIM	1.26	1.15	1.39	0.67

Table C.1 – Joint Coding Simulation Gain (IPPP GOP1; 2SRC)

Bit Rate	Group Name	Metric	mux_bit	mux_psnr	mux_ssim	mux_pspnr
256	AB	PSNR	0.31	0.21	0.22	0.18
256	AB	PSPNR	0.16	-0.13	-0.11	-0.21
256	AB	SSIM	2.68	2.66	2.65	2.73
256	AC	PSNR	-0.19	-0.38	-0.39	-0.39
256	AC	PSPNR	-0.6	-0.98	-0.93	-0.97
256	AC	SSIM	2.3	2.15	2.14	2.11
256	BC	PSNR	0.18	0.17	0.16	0.19
256	BC	PSPNR	0.21	0.15	0.17	0.19
256	BC	SSIM	0.02	-0.12	-0.07	-0.16
512	AB	PSNR	0.26	0.2	0.2	0.22
512	AB	PSPNR	0.11	-0.25	-0.21	-0.2
512	AB	SSIM	2.58	2.49	2.48	2.61
512	AC	PSNR	0.41	0.36	0.35	0.35
512	AC	PSPNR	-0.43	-0.77	-0.78	-0.81
512	AC	SSIM	6.06	6.48	6.43	6.49
512	BC	PSNR	0.26	0.27	0.28	0.26
512	BC	PSPNR	0.37	0.36	0.36	0.33
512	BC	SSIM	2.57	2.58	2.48	2.39

Table C.2 **– Joint Coding Simulation Gain (IPPP GOP2; 2SRC)**

Bit Rate	Group Name	Metric	mux_bit	mux_psnr	mux_ssim	mux_pspnr
256	AB	PSNR	2.02	1.36	1.33	0.56
256	AB	PSPNR	6.56	5.23	6.50	1.97
256	AB	SSIM	13.41	13.31	13.34	3.16
256	AC	PSNR	0.90	0.80	0.78	0.61
256	AC	PSPNR	2.44	1.84	2.16	1.44
256	AC	SSIM	10.43	9.38	10.14	5.34
256	BC	PSNR	1.73	1.73	1.71	0.15
256	BC	PSPNR	7.05	7.05	6.99	1.27
256	BC	SSIM	9.47	8.76	9.12	2.21
512	AB	PSNR	2.13	2.07	2.03	0.38
512	AB	PSPNR	3.57	1.61	1.44	1.60
512	AB	SSIM	7.57	6.13	6.08	2.24
512	AC	PSNR	0.61	0.36	0.53	0.65
512	AC	PSPNR	-0.88	-1.57	-0.08	-1.09
512	AC	SSIM	3.32	3.73	3.16	0.76
512	BC	PSNR	1.49	1.52	1.52	0.82
512	BC	PSPNR	7.12	7.25	7.25	-0.05
512	BC	SSIM	8.02	8.24	8.32	1.36

Table C.3 – Joint Coding Simulation Gain (IBBP GOP1; 2SRC)

Bit Rate	Group Name	Metric	mux_bit	mux_psnr	mux_ssim	mux_pspnr
256	AB	PSNR	0.22	0.13	0.16	0.02
256	AB	PSPNR	-0.48	-1.87	-1.81	-2.13
256	AB	SSIM	3.41	1.93	2.02	1.55
256	AC	PSNR	0.04	-0.60	-0.61	-0.64
256	AC	PSPNR	-4.60	-5.35	-5.67	-5.35
256	AC	SSIM	-1.13	-1.64	-2.05	-1.64
256	BC	PSNR	0.67	0.27	0.34	0.37
256	BC	PSPNR	-0.68	-1.30	-1.19	-1.15
256	BC	SSIM	1.96	3.75	3.32	3.40
512	AB	PSNR	1.04	0.68	0.76	0.59
512	AB	PSPNR	0.42	-0.89	-0.66	-1.15
512	AB	SSIM	3.82	3.32	3.40	3.11
512	AC	PSNR	-0.23	-1.11	-0.88	-1.10
512	AC	PSPNR	-4.18	-6.09	-5.65	-6.09
512	AC	SSIM	2.33	0.28	0.52	0.15
512	BC	PSNR	0.68	0.58	0.60	0.60
512	BC	PSPNR	0.13	-0.04	-0.01	-0.03
512	BC	SSIM	2.29	1.65	1.73	1.65

Table C.4 – Joint Coding Simulation Gain (IBBP GOP2; 2SRC)

Bit Rate	Group Name	Metric	mux_bit	mux_psnr	mux_ssim	mux_pspnr
256	AAB	PSNR	0.15	0.11	0.12	-0.36
256	AAB	PSPNR	-0.21	-0.25	-0.29	-1.75
256	AAB	SSIM	1.45	1.44	1.26	-2.32
256	ABB	PSNR	0.1	0.07	0.09	-0.31
256	ABB	PSPNR	-0.13	-0.24	-0.23	-0.84
256	ABB	SSIM	1.16	1.33	1.07	-0.62
256	ABC	PSNR	-0.18	-0.32	-0.25	-0.94
256	ABC	PSPNR	-0.76	-1.07	-0.9	-1.91
256	ABC	SSIM	2.17	1.84	2.02	1.61
256	ACC	PSNR	-0.23	-0.34	-0.27	-0.83
256	ACC	PSPNR	-0.78	-0.94	-0.88	-1.74
256	ACC	SSIM	-1.43	-1.77	-1.25	-3.18
256	BBC	PSNR	0.08	0.07	0.08	-0.01
256	BBC	PSPNR	0.06	0.04	0.06	-0.05
256	BBC	SSIM	0.67	0.65	0.74	0.21
256	BCC	PSNR	0.05	0.04	0.07	0.01
256	BCC	PSPNR	0.08	0.07	0.09	-0.05
256	BCC	SSIM	0.31	0.32	0.37	-1.18
512	AAB	PSNR	0.22	0.18	0.2	-0.21
512	AAB	PSPNR	-0.26	-0.35	-0.33	-1.66
512	AAB	SSIM	2.4	2.47	2.42	1.27
512	ABB	PSNR	0.17	0.13	0.14	-0.15
512	ABB	PSPNR	0.07	-0.1	-0.16	-0.64
512	ABB	SSIM	2.47	2.55	2.56	1.59
512	ABC	PSNR	0.32	0.12	0.27	-0.43
512	ABC	PSPNR	-0.33	-0.45	-0.64	-0.97
512	ABC	SSIM	2.87	2.78	3.02	2.22
512	ACC	PSNR	0.09	-0.03	0.12	0.37
512	ACC	PSPNR	-0.18	-0.21	-0.24	-1.97
512	ACC	SSIM	1.65	1.57	1.52	0.49
512	BBC	PSNR	0.13	0.11	0.12	0.03
512	BBC	PSPNR	0.14	0.1	0.13	0.03
512	BBC	SSIM	1.44	1.35	1.35	0.8
512	BCC	PSNR	0.09	0.08	0.09	0.01
512	BCC	PSPNR	0.06	0.03	0.07	-0.03
512	BCC	SSIM	0.47	0.41	0.49	0.27

Table C.5 – Joint Coding Simulation Gain (IPPP GOP1; 3SRC)

Bit Rate	Group Name	Metric	mux_bit	mux_psnr	mux_ssim	mux_pspnr
256	AAB	PSNR	0.29	0.13	0.17	0.11
256	AAB	PSPNR	0.2	-0.26	-0.18	-0.34
256	AAB	SSIM	2.91	2.44	2.55	2.38
256	ABB	PSNR	0.28	0.23	0.22	0.23
256	ABB	PSPNR	0.16	-0.02	-0.04	-0.05
256	ABB	SSIM	2.36	2.51	2.49	2.57
256	ABC	PSNR	0.17	0.04	0.04	0.03
256	ABC	PSPNR	-0.31	-0.63	-0.64	-0.67
256	ABC	SSIM	1.66	1.46	1.38	1.35
256	ACC	PSNR	0.18	0.11	0.13	0.08
256	ACC	PSPNR	-0.4	-0.6	-0.55	-0.62
256	ACC	SSIM	1.46	1.16	1.27	1.37
256	BBC	PSNR	0.16	0.13	0.15	0.15
256	BBC	PSPNR	0.19	0.13	0.17	0.16
256	BBC	SSIM	0.15	0.02	0.08	0.01
256	BCC	PSNR	0.16	0.14	0.17	0.09
256	BCC	PSPNR	0.16	0.15	0.17	0.07
256	BCC	SSIM	-0.09	-0.2	-0.05	-0.09
512	AAB	PSNR	0.27	0.18	0.18	0.17
512	AAB	PSPNR	0.2	-0.2	-0.23	-0.27
512	AAB	SSIM	2.77	2.56	2.55	2.48
512	ABB	PSNR	0.19	0.17	0.17	0.17
512	ABB	PSPNR	-0.01	-0.2	-0.19	-0.2
512	ABB	SSIM	1.94	2.11	2.03	2.1
512	ABC	PSNR	0.31	0.29	0.29	0.29
512	ABC	PSPNR	-0.22	-0.44	-0.43	-0.42
512	ABC	SSIM	4.36	4.74	4.75	4.79
512	ACC	PSNR	0.31	0.32	0.31	0.3
512	ACC	PSPNR	-0.39	-0.47	-0.44	-0.48
512	ACC	SSIM	4.7	4.94	4.94	4.91
512	BBC	PSNR	0.26	0.28	0.29	0.29
512	BBC	PSPNR	0.38	0.39	0.37	0.38
512	BBC	SSIM	2.47	2.64	2.65	2.67
512	BCC	PSNR	0.19	0.19	0.18	0.19
512	BCC	PSPNR	0.29	0.22	0.21	0.22
512	BCC	SSIM	1.73	1.92	1.9	1.82

Table C.6 – Joint Coding Simulation Gain (IPPP GOP2; 3SRC)

Bit Rate	Group Name	Metric	mux_bit	mux_psnr	mux_ssim	mux_pspnr
256	AAB	PSNR	1.33	1.27	1.17	0.21
256	AAB	PSPNR	4.13	5.49	5.13	1.80
256	AAB	SSIM	14.06	15.56	15.44	6.00
256	ABB	PSNR	2.16	2.85	2.44	1.41
256	ABB	PSPNR	7.64	7.66	7.69	1.65
256	ABB	SSIM	16.02	16.50	16.32	4.72
256	ABC	PSNR	2.19	1.71	1.48	0.03
256	ABC	PSPNR	7.41	6.48	7.36	3.48
256	ABC	SSIM	14.05	14.62	13.91	4.52
256	ACC	PSNR	2.21	2.35	2.40	2.59
256	ACC	PSPNR	8.45	8.78	8.90	3.55
256	ACC	SSIM	21.29	21.69	22.48	6.54
256	BBC	PSNR	2.13	2.01	2.12	1.00
256	BBC	PSPNR	17.09	16.89	17.05	0.89
256	BBC	SSIM	6.38	6.29	6.67	1.33
256	BCC	PSNR	1.44	1.39	1.46	1.87
256	BCC	PSPNR	11.92	11.81	11.94	1.24
256	BCC	SSIM	6.48	6.10	6.52	0.06
512	AAB	PSNR	1.59	2.05	1.86	0.59
512	AAB	PSPNR	9.18	4.61	4.40	0.83
512	AAB	SSIM	5.10	5.75	5.34	1.84
512	ABB	PSNR	2.70	2.65	2.65	1.26
512	ABB	PSPNR	3.75	3.75	3.66	0.54
512	ABB	SSIM	7.27	10.12	10.15	1.64
512	ABC	PSNR	3.61	3.28	3.26	1.33
512	ABC	PSPNR	2.09	1.84	1.81	0.66
512	ABC	SSIM	10.79	9.87	11.40	2.94
512	ACC	PSNR	0.93	0.66	0.70	0.29
512	ACC	PSPNR	0.88	0.72	0.59	0.66
512	ACC	SSIM	16.13	16.98	18.63	2.76
512	BBC	PSNR	1.13	1.10	1.03	0.81
512	BBC	PSPNR	3.94	3.91	3.88	3.12
512	BBC	SSIM	21.94	21.19	22.08	7.89
512	BCC	PSNR	1.01	0.81	0.98	0.55
512	BCC	PSPNR	3.30	3.23	3.31	1.70
512	BCC	SSIM	3.07	3.37	4.30	3.31

Table C.7 – Joint Coding Simulation Gain (IBBP GOP1; 3SRC)

Bit Rate	Group Name	Metric	mux_bit	mux_psnr	mux_ssim	mux_pspnr
256	AAB	PSNR	1.17	1.13	1.16	0.48
256	AAB	PSPNR	0.47	-1.26	-1.28	-1.28
256	AAB	SSIM	2.84	1.27	1.37	1.21
256	ABB	PSNR	0.87	0.91	0.89	0.39
256	ABB	PSPNR	-0.03	-1.16	-1.06	-1.15
256	ABB	SSIM	3.69	2.25	2.43	1.29
256	ABC	PSNR	0.45	0.43	0.48	0.16
256	ABC	PSPNR	-3.13	-4.09	-4.28	-4.19
256	ABC	SSIM	2.31	4.52	5.07	1.61
256	ACC	PSNR	-0.57	-0.69	-0.65	-0.64
256	ACC	PSPNR	-3.16	-3.42	-3.35	-3.34
256	ACC	SSIM	-1.41	-1.77	-1.56	-1.55
256	BBC	PSNR	1.13	1.00	0.97	-0.01
256	BBC	PSPNR	-0.58	-1.16	-0.96	-0.97
256	BBC	SSIM	1.68	2.97	2.31	1.35
256	BCC	PSNR	0.60	0.58	0.59	-0.05
256	BCC	PSPNR	-0.44	-1.02	-0.96	-0.97
256	BCC	SSIM	-0.82	-1.63	-1.53	-1.55
512	AAB	PSNR	0.88	0.86	0.83	0.60
512	AAB	PSPNR	0.18	-2.14	-1.11	-1.19
512	AAB	SSIM	2.75	1.42	2.40	1.65
512	ABB	PSNR	1.11	0.97	0.90	0.72
512	ABB	PSPNR	0.58	-0.59	-0.09	-1.27
512	ABB	SSIM	4.26	3.40	3.96	1.77
512	ABC	PSNR	0.01	0.82	-0.37	-0.51
512	ABC	PSPNR	-2.85	-5.03	-3.72	-3.36
512	ABC	SSIM	1.99	1.56	1.02	0.63
512	ACC	PSNR	-0.35	-0.11	-0.68	-0.67
512	ACC	PSPNR	-3.76	-5.92	-4.3	-3.55
512	ACC	SSIM	0.46	0.51	0.14	-0.36
512	BBC	PSNR	0.51	0.04	0.48	-0.05
512	BBC	PSPNR	-0.2	-1.05	-0.23	-0.98
512	BBC	SSIM	1.31	0.82	1.21	-0.60
512	BCC	PSNR	0.39	0.28	0.58	0.08
512	BCC	PSPNR	-0.34	-0.53	-0.01	-1.05
512	BCC	SSIM	1.27	1.02	2.65	0.07

Table C.8 – Joint Coding Simulation Gain (IBBP GOP2; 3SRC)

	Mux Bit			Mux PSNR			Mux SSIM			Mux PSPNR		
	PSNR	PSPNR	SSIM	PSNR	PSPNR	SSIM	PSNR	PSPNR	SSIM	PSNR	PSPNR	SSIM
Aki	40.59	57.98	77.65	41.28	60.06	80.50	40.18	59.03	75.35	40.77	59.72	77.21
Fot	29.09	34.00	17.03	30.72	35.40	21.12	30.95	38.06	29.24	30.50	36.32	23.93
Hal	37.95	48.35	63.62	38.16	50.28	67.23	37.94	50.16	66.78	38.04	50.11	67.04
mad	37.72	51.78	70.13	38.73	53.06	74.02	38.72	52.36	72.57	38.75	52.94	73.96
Mcl	30.68	38.19	61.24	27.75	32.79	44.98	28.21	33.23	46.40	28.20	33.57	47.51
Sil	35.95	45.24	60.06	36.15	45.50	60.95	36.09	45.24	60.97	36.30	45.60	61.15

Table C.9 – Picture Quality Results for 6SRC (IBBP GOP1; 256 kbps)

	Mux Bit			Mux PSNR			Mux SSIM			Mux PSPNR		
	PSNR	PSPNR	SSIM	PSNR	PSPNR	SSIM	PSNR	PSPNR	SSIM	PSNR	PSPNR	SSIM
Aki	41.46	60.31	84.92	43.61	66.40	86.72	43.61	66.41	86.73	42.97	63.88	85.71
Fot	31.58	38.97	27.56	32.97	41.42	36.42	33.51	42.28	44.09	33.62	41.11	40.53
Hal	38.74	52.52	69.32	39.36	53.78	70.91	39.40	53.98	71.51	39.42	54.02	71.06
mad	39.76	54.49	77.15	40.91	58.99	78.60	40.92	58.90	77.36	40.79	58.88	79.29
mcl	33.84	42.28	71.23	31.21	37.05	61.91	29.91	36.01	60.51	30.44	37.45	62.23
sil	39.87	54.02	77.10	39.71	55.01	75.24	39.55	53.00	75.27	39.86	55.02	75.58

Table C.10 – Picture Quality Results for 6SRC (IBBP GOP1; 512kbps)

	Mux Bit			Mux PSNR			Mux SSIM			Mux PSPNR		
	PSNR	PSPNR	SSIM	PSNR	PSPNR	SSIM	PSNR	PSPNR	SSIM	PSNR	PSPNR	SSIM
aki	34.25	44.88	66.93	37.12	46.47	70.51	36.33	47.21	69.70	36.74	45.16	70.15
fot	28.42	33.62	13.99	29.25	34.06	15.46	29.51	34.43	16.52	29.10	33.86	14.89
hal	30.40	38.28	50.55	33.03	40.07	52.19	32.90	39.53	51.64	32.88	39.86	52.04
mad	32.44	39.66	45.79	34.14	42.41	53.43	34.09	41.52	52.69	33.94	41.65	52.79
mcl	24.30	28.96	21.81	21.90	24.45	8.37	21.77	24.26	7.63	22.42	25.81	7.20
sil	30.86	35.76	25.75	30.93	35.90	26.12	31.53	35.77	30.50	30.86	36.02	28.83

Table C.11 – Picture Quality Results for 6SRC (IPPP GOP1; 256 kbps)

	Mux Bit			Mux PSNR			Mux SSIM			Mux PSPNR		
	PSNR	PSPNR	SSIM	PSNR	PSPNR	SSIM	PSNR	PSPNR	SSIM	PSNR	PSPNR	SSIM
aki	37.76	47.57	71.55	40.81	56.92	80.63	40.52	56.56	79.78	40.05	53.73	78.69
fot	30.73	37.15	24.98	31.63	38.31	29.53	31.82	38.61	30.79	31.79	38.39	26.58
hal	34.67	43.53	58.72	36.00	47.24	62.97	36.07	46.95	61.23	35.91	46.71	60.79
mad	35.76	44.99	60.80	37.39	49.29	68.70	37.40	49.05	68.91	37.17	48.56	66.17
mcl	27.59	34.38	40.30	24.78	28.44	19.77	24.54	28.34	21.99	25.12	29.27	21.97
sil	32.95	39.65	42.35	33.86	40.77	45.66	34.12	41.16	52.97	34.08	40.86	44.29

Table C.12 – Picture Quality Results for 6SRC (IPPP GOP1; 512kbps)

C.3 SAMVIQ Sessions Results

HRC	a1	a2	a3	a4	a5	a6	a7	a8	a9	a10	a11	a12	a13	a14	a15
HRC1	52	54	60	75	67	67	71	53	69	62	61	55	59	63	67
HRC2	66	62	69	80	77	76	78	62	77	73	68	63	67	74	76
HRC3	47	45	49	68	57	61	64	47	69	64	62	57	55	58	70
HRC4	59	55	60	77	77	77	74	58	81	73	72	67	64	68	77
HRC5	50	55	61	74	58	66	65	51	73	61	64	57	62	61	72
HRC6	62	61	73	79	74	73	74	62	81	75	78	68	65	71	78
HRC7	51	51	65	78	62	69	67	60	69	61	62	57	58	68	73
HRC8	62	63	77	83	71	76	75	66	77	76	72	67	75	76	83
HRC9	50	51	58	68	61	64	64	52	66	67	55	61	54	66	73
HRC10	62	59	73	80	71	74	72	61	77	73	69	68	66	73	83
SRC1	71	84	90	94	89	93	85	80	90	87	85	84	91	86	91
SRC2	36	29	40	55	30	44	42	20	47	38	49	26	20	26	38
SRC3	56	64	66	74	70	77	74	59	71	74	60	63	71	68	78
SRC4	64	68	72	85	85	78	84	74	84	73	73	81	74	88	85
SRC5	60	45	55	81	71	61	68	55	80	71	61	58	47	80	82
SRC6	49	43	62	67	60	69	69	56	71	68	71	60	72	60	77

Table C.13 – MOS for SRC and HRC per Observer (IBBP GOP1)

HRC	a1	a2	a3	a4	a5	a6	a7	a8	a9	a10	a11	a12	a13	a14	a15
HRC1	26	25	40	51	41	41	51	30	50	44	43	35	37	39	47
HRC2	47	42	54	67	59	58	62	44	68	58	64	54	58	58	68
HRC3	29	31	38	50	41	41	44	31	46	40	41	41	42	41	52
HRC4	46	39	48	59	56	59	59	46	62	53	55	51	54	54	65
HRC5	28	29	40	47	40	45	42	33	58	43	42	38	44	40	56
HRC6	46	43	61	67	65	56	62	49	68	59	65	57	56	57	70
HRC7	34	29	49	55	44	44	49	36	48	46	37	33	34	39	51
HRC8	47	43	63	70	60	61	64	48	68	60	67	59	60	62	73
HRC9	29	32	34	42	38	40	46	35	48	41	37	34	31	40	57
HRC10	44	44	57	61	59	54	62	45	64	53	61	54	62	53	76
SRC1	56	63	74	80	82	78	79	64	77	72	70	68	75	72	82
SRC2	29	14	39	40	27	22	34	16	41	35	35	21	22	19	36
SRC3	41	54	52	57	53	59	58	44	58	54	53	54	60	53	68
SRC4	51	53	55	74	71	64	70	59	74	58	62	67	53	70	73
SRC5	20	13	21	39	29	24	28	18	41	28	27	21	19	35	49
SRC6	29	16	49	51	41	52	55	37	57	52	62	42	57	41	62

Table C.14 – MOS for SRC and HRC per Observer (IPPP GOP1)

HRC	a1	a2	a3	a4	a5	a6	a7	a8	a9	a10	a11	a12	a13	a14	a15
HRC1	0.58	0.57	0.63	0.76	0.69	0.68	0.75	0.56	0.72	0.67	0.62	0.59	0.61	0.65	0.71
HRC2	0.73	0.65	0.72	0.81	0.80	0.77	0.82	0.65	0.80	0.79	0.69	0.68	0.70	0.77	0.81
HRC3	0.52	0.47	0.51	0.68	0.60	0.62	0.67	0.50	0.73	0.70	0.62	0.62	0.57	0.60	0.74
HRC4	0.65	0.58	0.64	0.78	0.80	0.79	0.78	0.62	0.85	0.79	0.73	0.72	0.66	0.71	0.81
HRC5	0.55	0.58	0.65	0.75	0.60	0.68	0.68	0.54	0.76	0.66	0.65	0.61	0.65	0.63	0.76
HRC6	0.69	0.63	0.77	0.80	0.77	0.74	0.79	0.65	0.85	0.81	0.79	0.74	0.68	0.74	0.82
HRC7	0.56	0.54	0.69	0.79	0.64	0.70	0.70	0.64	0.72	0.66	0.63	0.62	0.61	0.71	0.77
HRC8	0.68	0.66	0.82	0.84	0.73	0.78	0.79	0.70	0.81	0.82	0.73	0.73	0.79	0.80	0.88
HRC9	0.56	0.54	0.61	0.69	0.63	0.66	0.68	0.55	0.69	0.73	0.56	0.66	0.56	0.69	0.77
HRC10	0.68	0.62	0.77	0.82	0.74	0.76	0.76	0.64	0.81	0.79	0.70	0.74	0.69	0.75	0.88
SRC1	0.79	0.88	0.90	0.95	0.89	0.93	0.90	0.80	0.90	0.92	0.86	0.89	0.91	0.86	0.95
SRC2	0.43	0.32	0.43	0.61	0.32	0.44	0.44	0.21	0.55	0.42	0.51	0.29	0.25	0.29	0.44
SRC3	0.64	0.71	0.71	0.74	0.74	0.79	0.77	0.65	0.75	0.82	0.60	0.68	0.71	0.72	0.82
SRC4	0.72	0.72	0.76	0.85	0.85	0.83	0.89	0.78	0.84	0.77	0.73	0.85	0.80	0.91	0.89
SRC5	0.63	0.45	0.58	0.81	0.75	0.63	0.68	0.60	0.85	0.78	0.61	0.62	0.51	0.80	0.82
SRC6	0.52	0.43	0.69	0.67	0.65	0.69	0.77	0.59	0.75	0.75	0.71	0.70	0.74	0.65	0.83

Table C.15 – Normalised MOS for SRC and HRC per Observer (IBBP GOP1)

HRC	a1	a2	a3	a4	a5	a6	a7	a8	a9	a10	a11	a12	a13	a14	a15
HRC1	0.27	0.26	0.41	0.52	0.42	0.44	0.53	0.31	0.52	0.50	0.44	0.38	0.38	0.41	0.48
HRC2	0.50	0.44	0.56	0.69	0.60	0.61	0.65	0.46	0.71	0.65	0.66	0.59	0.61	0.60	0.70
HRC3	0.30	0.32	0.40	0.51	0.42	0.44	0.45	0.32	0.48	0.44	0.42	0.45	0.44	0.43	0.54
HRC4	0.49	0.41	0.50	0.61	0.57	0.62	0.60	0.49	0.64	0.60	0.56	0.55	0.56	0.57	0.67
HRC5	0.29	0.30	0.41	0.48	0.41	0.48	0.43	0.34	0.61	0.49	0.44	0.41	0.45	0.41	0.58
HRC6	0.49	0.45	0.63	0.69	0.67	0.59	0.65	0.51	0.71	0.66	0.67	0.61	0.59	0.60	0.72
HRC7	0.37	0.31	0.50	0.56	0.45	0.47	0.50	0.37	0.50	0.52	0.38	0.35	0.35	0.41	0.52
HRC8	0.50	0.46	0.65	0.72	0.62	0.65	0.66	0.50	0.71	0.68	0.69	0.64	0.63	0.66	0.75
HRC9	0.31	0.34	0.36	0.42	0.39	0.42	0.47	0.37	0.49	0.46	0.38	0.37	0.32	0.42	0.59
HRC10	0.46	0.46	0.59	0.62	0.61	0.58	0.64	0.47	0.67	0.59	0.63	0.59	0.64	0.56	0.78
SRC1	0.56	0.64	0.74	0.80	0.82	0.80	0.81	0.66	0.77	0.75	0.71	0.68	0.75	0.73	0.82
SRC2	0.34	0.18	0.43	0.45	0.30	0.23	0.38	0.17	0.47	0.41	0.37	0.23	0.27	0.22	0.40
SRC3	0.44	0.54	0.56	0.60	0.53	0.64	0.63	0.44	0.58	0.62	0.53	0.58	0.63	0.54	0.68
SRC4	0.55	0.59	0.56	0.74	0.73	0.75	0.70	0.62	0.82	0.68	0.67	0.80	0.56	0.78	0.80
SRC5	0.20	0.14	0.21	0.39	0.29	0.24	0.28	0.19	0.41	0.31	0.27	0.22	0.20	0.35	0.49
SRC6	0.30	0.16	0.51	0.51	0.43	0.52	0.55	0.39	0.57	0.57	0.62	0.45	0.57	0.43	0.62

Table C.16 – Normalised MOS for SRC and HRC per Observer (IPPP GOP1)

SRC	SRC1 (Akiyo)			SRC2 (Fot)			SRC3 (Hall)			SRC4 (MAD)			SRC5 (MCL)			SRC6 (SIL)		
HRC	μ	CI	σ	μ	CI	σ	μ	CI	σ	μ	CI	σ	μ	CI	σ	μ	CI	σ
HRC1	73.1	4.7	9.4	16.3	7.6	15.0	45.7	5.5	10.8	57.9	8.2	16.2	8.5	3.5	6.8	38.2	7.0	13.9
HRC2	81.9	3.5	7.0	37.2	6.5	12.8	69.5	5.9	11.7	75.1	5.5	10.8	23.9	8.5	16.9	56.1	8.3	16.3
HRC3	66.1	4.5	8.9	11.3	9.3	18.4	43.1	5.1	10.1	51.2	6.0	11.9	34.8	3.8	7.4	34.1	6.4	12.6
HRC4	70.8	4.7	9.4	33.0	5.9	11.6	56.9	4.1	8.0	66.1	4.4	8.7	47.8	7.7	15.3	47.3	8.1	16.0
HRC5	68.8	6.0	11.9	17.5	8.6	17.0	48.9	4.3	8.4	58.9	6.0	11.9	18.3	5.0	9.9	36.9	7.9	15.6
HRC6	77.8	6.2	12.2	46.0	5.9	11.7	65.1	4.5	8.9	71.3	4.4	8.6	35.5	7.8	15.5	56.9	8.1	16.1
HRC7	68.5	6.0	11.9	21.6	4.3	8.5	44.5	5.4	10.6	59.6	5.6	11.1	16.5	5.9	11.7	40.9	7.8	15.3
HRC8	79.5	5.1	10.0	48.2	5.3	10.4	64.4	4.8	9.5	69.7	4.5	8.9	36.4	8.5	16.8	63.9	10.0	19.8
HRC9	66.8	5.6	11.0	15.1	7.9	15.5	46.0	5.5	11.0	58.3	5.4	10.6	12.6	4.8	9.6	34.3	8.5	16.7
HRC10	74.5	5.3	10.4	36.1	6.3	12.5	60.6	4.5	8.8	67.5	4.7	9.2	40.3	7.7	15.2	60.1	8.4	16.6

Table C.17 – Mean (μ), Confidence of Interval at 95% (CI) and Standard Deviation (σ) of all SRCs per HRC (IPPP GOP1)

SRC	SRC1 (Akiyo)			SRC2 (Fot)			SRC3 (Hall)			SRC4 (MAD)			SRC5 (MCL)			SRC6 (SIL)		
HRC	μ	CI	σ	μ	CI	σ	μ	CI	σ	μ	CI	σ	μ	CI	σ	μ	CI	σ
HRC1	87.3	4.3	8.5	25.3	9.0	17.8	66.6	6.1	12.0	79.1	4.7	9.2	56.5	10.2	20.1	58.1	5.1	10.1
HRC2	93.7	2.9	5.8	40.1	7.6	15.1	74.9	4.7	9.2	83.9	3.5	6.9	64.0	9.2	18.2	69.7	4.8	9.6
HRC3	78.1	4.5	8.9	22.0	8.0	15.8	59.9	6.0	11.8	66.2	4.0	7.9	67.7	12.0	23.7	56.2	7.2	14.3
HRC4	86.3	4.1	8.0	33.9	8.2	16.2	69.8	6.2	12.3	81.4	5.7	11.3	75.1	9.8	19.4	69.6	6.3	12.4
HRC5	84.5	5.0	9.8	27.7	8.0	15.8	65.5	5.0	9.9	74.2	6.4	12.7	58.3	8.8	17.4	59.6	6.6	13.0
HRC6	91.2	3.9	7.7	43.1	7.4	14.5	72.6	5.2	10.2	83.3	4.8	9.5	69.1	8.7	17.2	70.2	5.8	11.5
HRC7	82.6	5.1	10.0	37.7	6.9	13.6	66.0	7.0	13.8	74.3	6.7	13.2	60.7	8.0	15.8	57.8	5.3	10.6
HRC8	92.9	2.9	5.7	52.8	6.1	12.1	75.1	3.9	7.6	82.1	4.1	8.0	66.6	7.4	14.6	70.8	4.4	8.6
HRC9	81.3	4.6	9.1	32.1	7.9	15.6	63.3	7.4	14.6	71.6	6.7	13.2	59.6	10.1	20.0	56.0	6.8	13.4
HRC10	87.7	4.0	7.8	44.8	7.1	14.0	70.8	6.6	13.1	82.9	5.8	11.4	70.5	9.0	17.8	67.5	6.2	12.3

Table C.18 – Mean (μ), Confidence of Interval at 95% (CI) and Standard Deviation (σ) of all SRCs per HRC (IBBP GOP1)

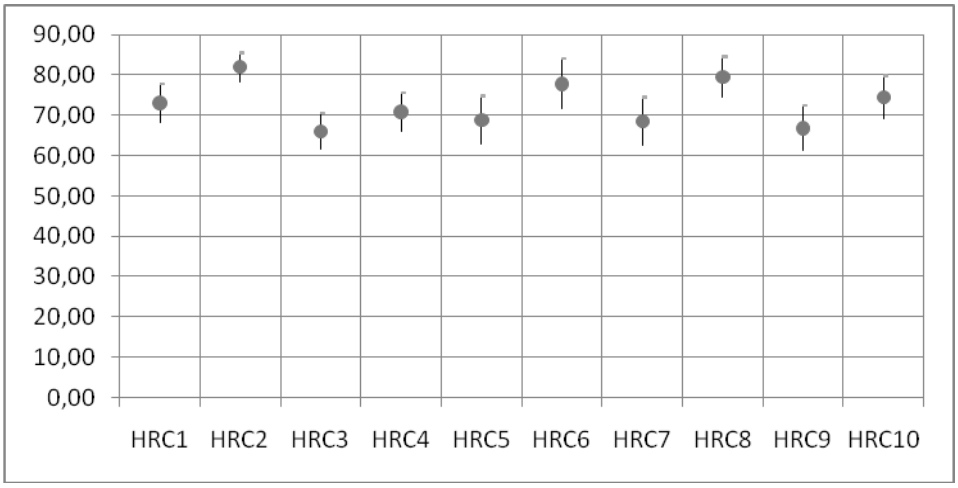


Figure C.9 – MOS values and 95% CI for the content Akiyo (IPPP GOP1)

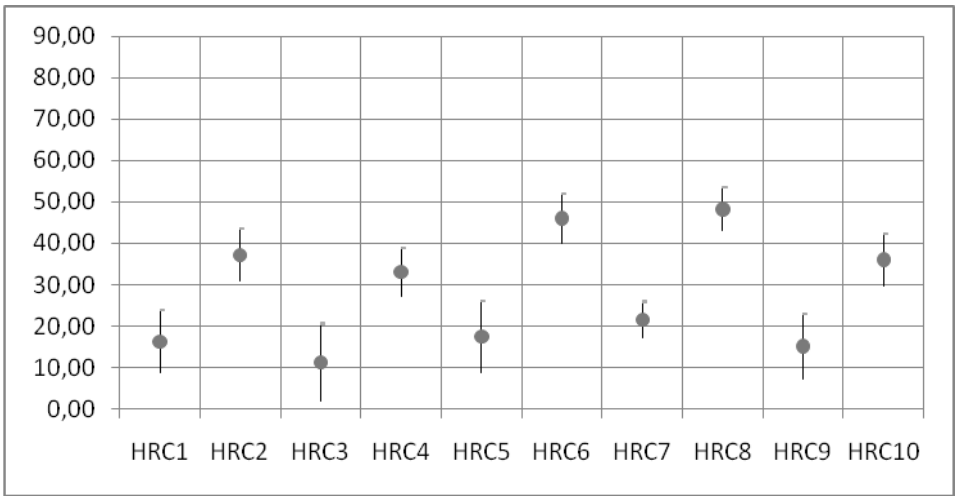


Figure C.10 – MOS values and 95% CI for the content Fot (IPPP GOP1)

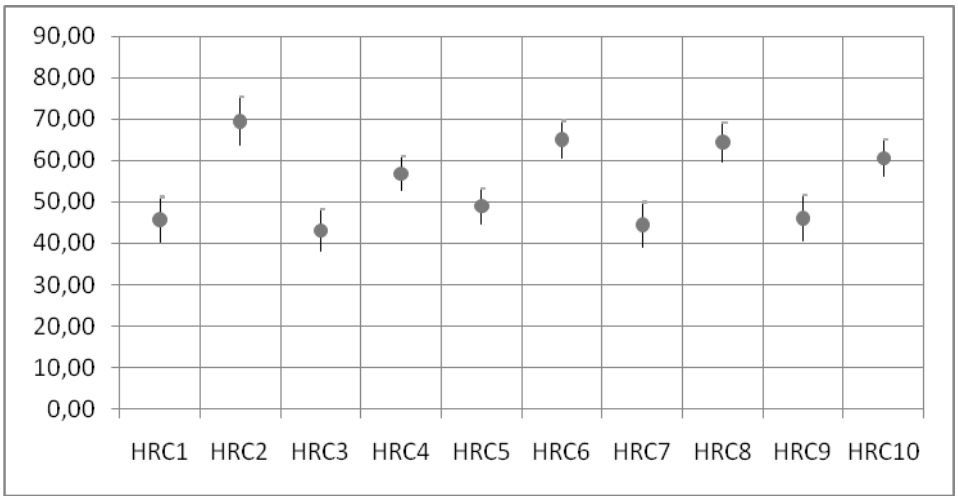


Figure C.11 – MOS values and 95% CI for the content Hall (IPPP GOP1)

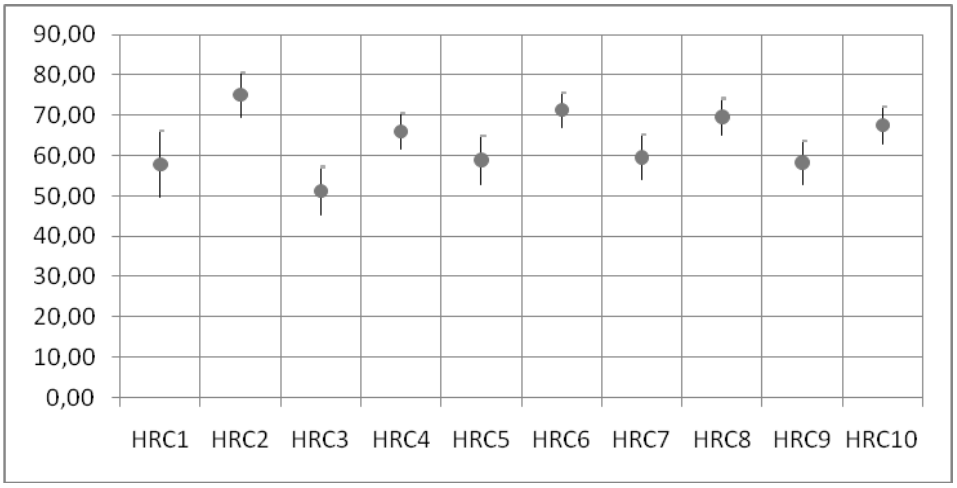


Figure C.12 – MOS values and 95% CI for the content MAD (IPPP GOP1)

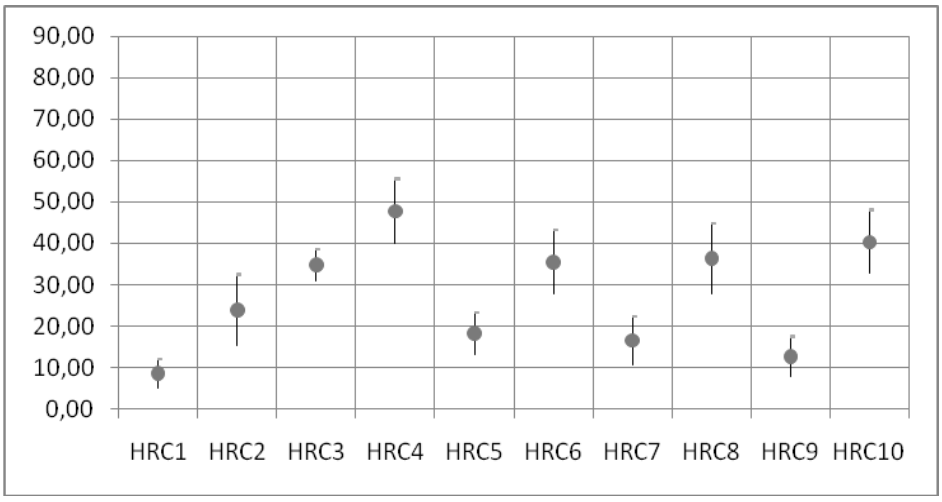


Figure C.13 – MOS values and 95% CI for the content MCL (IPPP GOP1)

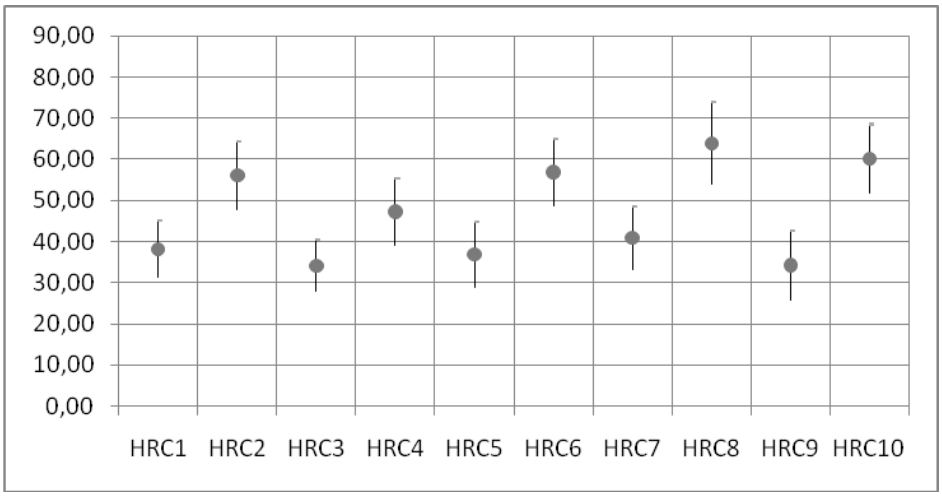


Figure C.14 – MOS values and 95% CI for the content SIL (IPPP GOP1)

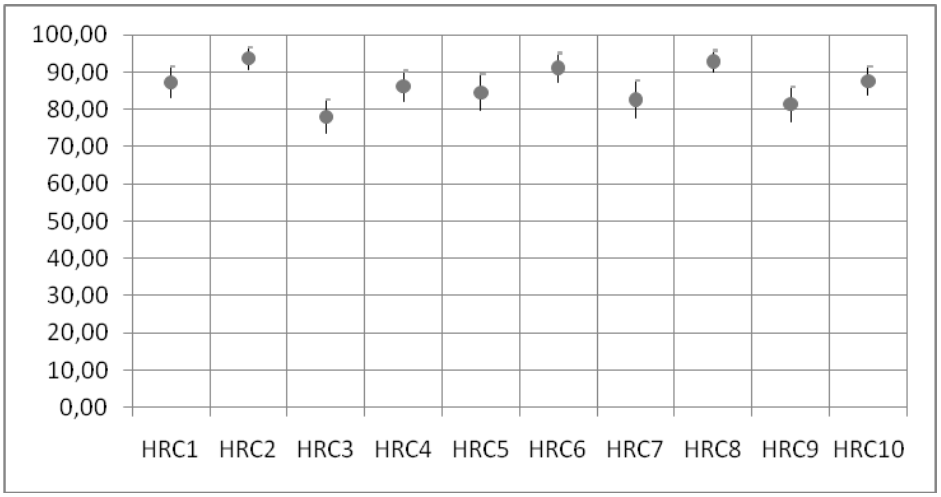


Figure C.15 – MOS values and 95% CI for the content Akiyo (IBBP GOP1)

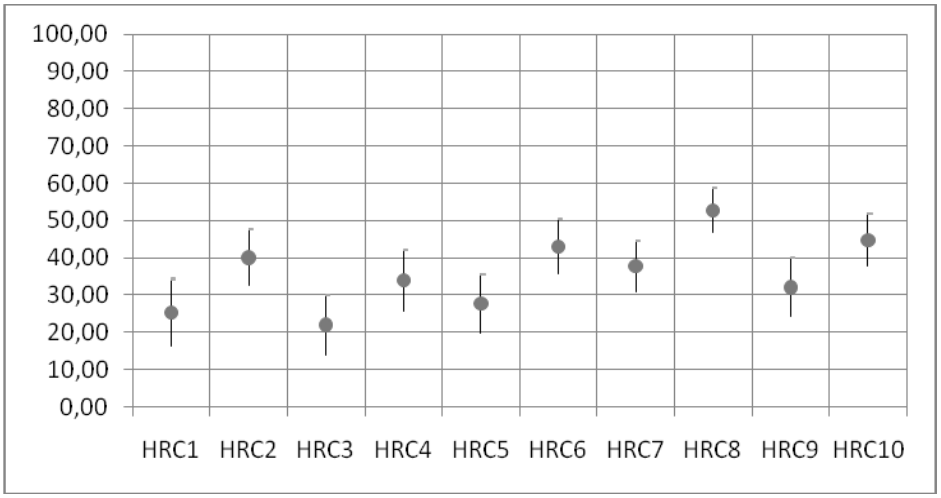


Figure C.16 – MOS values and 95% CI for the content Fot (IBBP GOP1)

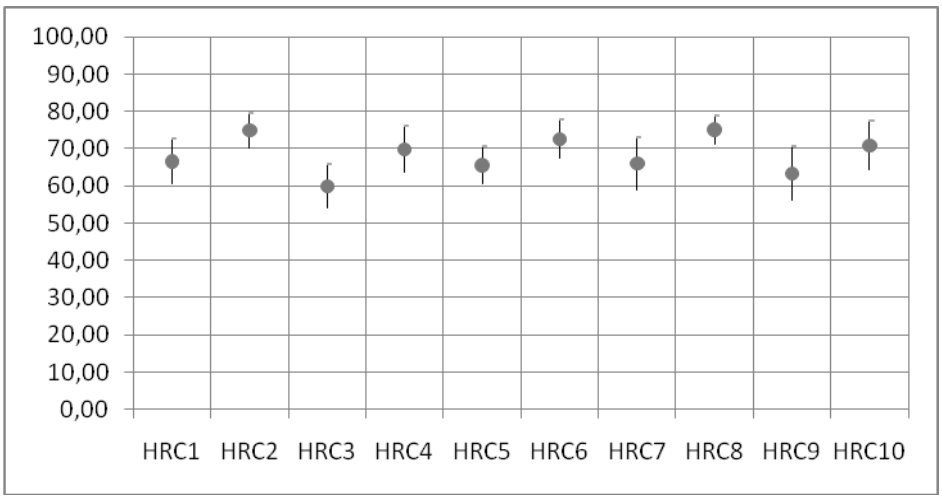


Figure C.17 – MOS values and 95% CI for the content Hall (IBBP GOP1)

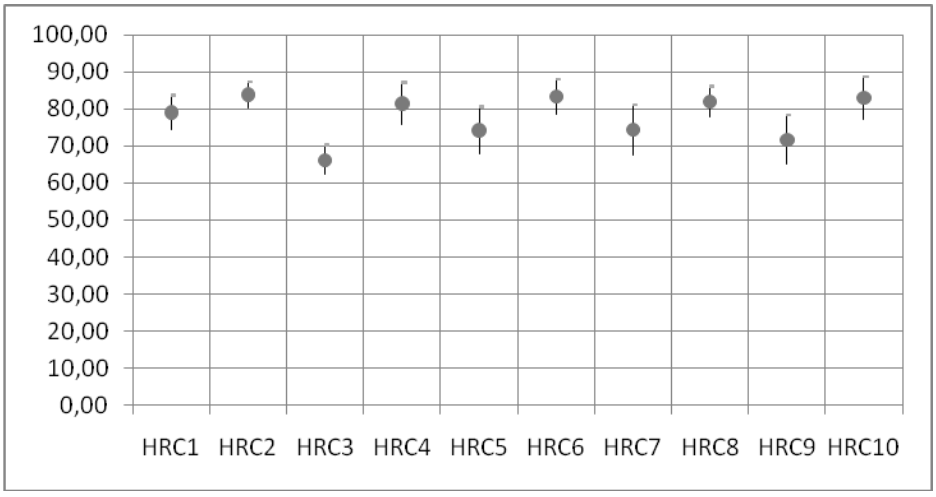


Figure C.18 – MOS values and 95% CI for the content MADo (IBBP GOP1)

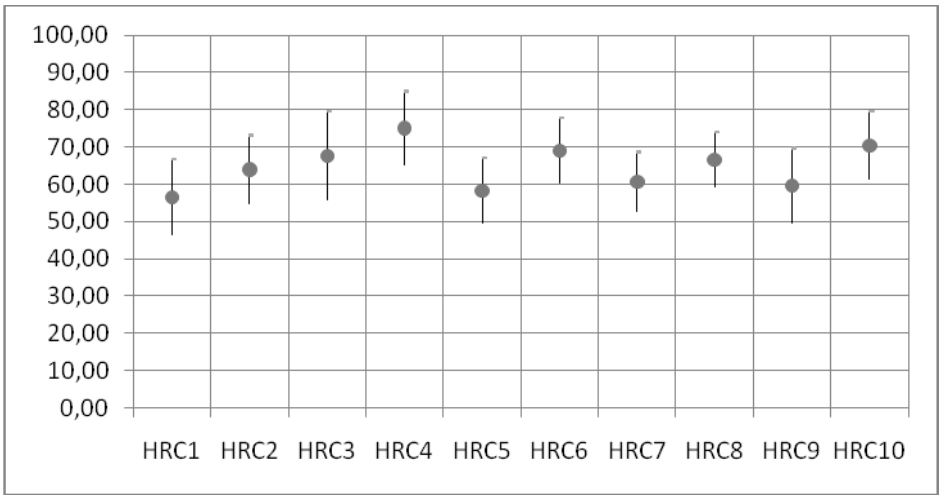


Figure C.19 – MOS values and 95% CI for the content MCL (IBBP GOP1)

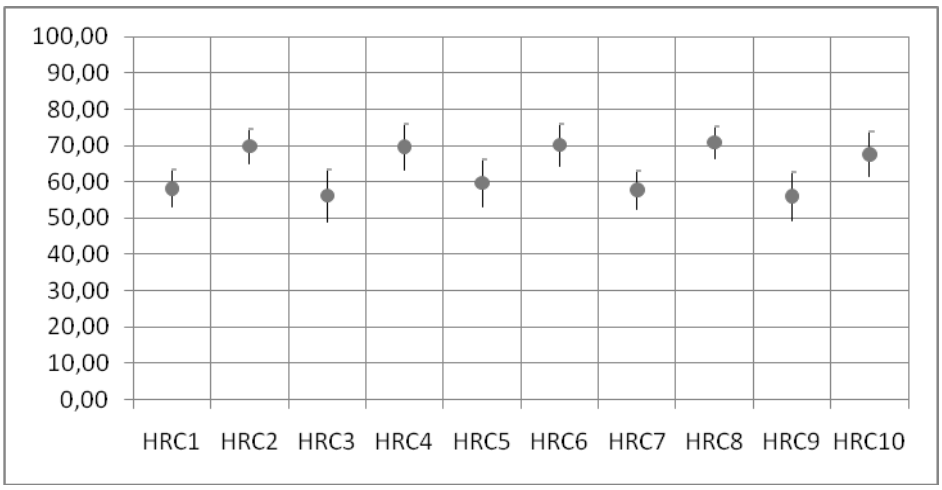


Figure C.20 – MOS values and 95% CI for the content SIL (IBBP GOP1)

References

[1] YouTube, “YouTube Statistics.” Available from: http://www.youtube.com/t/press_statistics, Retrieved 31 January 2012.

[2] ISO/IEC IS 13818-2 | ITU-T Recommendation H.262, “Information Technology - Generic Coding of Moving Pictures and Associated Audio - Part 2: Video”, 1994.

[3] Ulrich Reimers (Editor), “Digital Video Broadcasting (DVB) – The International Standard for Digital Television (2nd Edition)”, Springer, ISBN 354043545X, 2004.

[4] Digital Video Broadcasting (DVB): Transmission Systems for Handheld Terminals, ETSI standard, EN 302 304 V1.1.1, ETSI, 2004.

[5] Gerard Faria, Jukka A. Henriksson, Erik Stare, and Pekka Talmola, “DVB-H: Digital broadcast services to handheld devices”, Proceedings of the IEEE, Volume 94, Issue 1, Pages 194-209, January 2006.

[6] ITU-T and ISO/IEC JTC 1, “Advanced Video Coding for Generic Audiovisual Services”, ITU-T Rec. H.264 & ISO/IEC 14496-10, (6th Ed.). Available from: http://www.itu.int/rec/T-REC-H.264.

[7] Joern Ostermann, Jan Bormans, Peter List, Detlev Marpe, Matthias Narroschke, Fernando Pereira, Thomas Stockhammer, and Thomas Wedi, “Video coding with H.264/AVC: Tools, Performance, and Complexity”, IEEE Circuits and Systems Magazine, Volume 4, Issue 1, Pages 7-28, First Quarter 2004.

[8] Ajay Luthra, Garry J. Sullivan, and Thomas Wiegand (Editors), IEEE Transactions on Circuits Systems for Video Technology (Special Issue on the H.264/AVC Video Coding Standard), Volume 13, Issue 7, July 2003.

[9] Thomas Wiegand, and Gary J. Sullivan, “The H.264/AVC Video Coding Standard {Standards in a nutshell}”, IEEE Signal Processing Magazine, Volume 24, Issue 2, March 2007.

[10] Michael Kornfeld, and Gunther May, “DVB-H and IP Datacast—Broadcast to Handheld Devices”, IEEE Transactions on Broadcasting, Volume 53, Issue 1, Pages 161-170, March 2007.

[11] MPEG, “Vision, applications and requirements for high efficiency video coding (HEVC)”, ISO/IEC/JTC1/SC29/WG11 N11872, Daegu, South Korea, January 2011.

[12] H265.net. Available from: http://www.h265.net/, Retrieved 31 January 2012.

[13] Thomas Wiegand, Jens-Rainer Ohm, Garry J. Sullivan, Woo-Jin. Han, Rajan Joshi, Thiow Keng Tan, and Kemal Ugur, “Special section on the joint call for proposals on high efficiency video coding (HEVC) standardisation”, IEEE Transactions on Circuits System for Video Technology, Volume 20, Issue 12, Pages 1661-1666, December 2010.

[14] Kemal Ugur, Kenneth Andersson, Arild Fuldseth, Gisle Bjøntegaard, Lars Petter Endresen, Jani Lainema, Antti Hallapuro, Justin Ridge, Dmytro Rusanovskyy, Cixun Zhang, Andrey Norkin, Clinton Priddle, Thomas Rusert, Jonatan Samuelsson, Rickard Sjöberg, and Zhuangfei Wu, “High performance, low complexity video coding and the emerging HEVC standard”, IEEE Transactions on Circuits System for Video Technology, Volume 20, Issue 12, Pages 1688-1697, December 2010.

[15] Kannan Ramchandran, Antonio Ortega, and Martin Vetterli, “Bit allocation for dependent quantisation with applications to multiresolution and MPEG video coders”, IEEE Transactions on Image Processing, Volume 3, Issue 5, Pages 533–545, 1994.

[16] H. Everett, “Generalized Lagrange multiplier method for solving problems of optimum allocation of resource”, Operations Research, Volume 11, Pages 399–417, 1963.

[17] Richard E. Bellman, “Dynamic Programming”, Princeton University Press, ISBN 9780691146683, 2010.

[18] Antonio Ortega, “Optimal bit allocation under multiple rate constraints”, Proceedings of the 1996 Data Compression Conference (DCC 1996), Snowbird, U.S.A., Pages 349–358, 1996.

[19] MPEG Test Model Editing Committee, “MPEG Video Test Model 5 (TM-5)”, ISO/IEC-JTC1/SC29/WG11, Document N0400, Sydney MPEG meeting April 1993.

[20] ITU-T SG16/Q15, “Video Codec Test Model, Near-Term, Version 8 (TMN8)”, Document Q15-A-59, Portland, U.S.A., June 1997.

[21] Jordi Ribas-Corbera, and Shawmin Lei, “Rate control in DCT video coding for low-delay communications”, IEEE Transactions on Circuits and Systems for Video Technology, Volume 9, Issue 1, Pages 172-185, February 1999.

[22] Hsueh-Ming Hang, and Jiann-Jone Chen, “Source model for transform video coder and its application – part I: fundamental theory”, IEEE Transactions Circuits and Systems for Video Technology, Volume 7, Issue 2, Pages 287-298, April 1997.

[23] Tihao Chiang, and Ya-Qin Zhang, “A new rate control scheme using quadratic rate distortion model”, IEEE Transactions on Circuits Systems for Video Technology, Volume 7, Issue 1, Pages 246-250, February 1997.

[24] Wei Ding, and Bede Liu, “Rate control of MPEG video coding and recording by rate-quantisation modeling”, IEEE Transactions on Circuits Systems for Video Technology, Volume 6, Issue 1, Pages 12-20, February 1996.

[25] Liang-jin Lin, and A. Ortega, “Bit-rate control using piecewise approximated rate-distortion characteristics”, IEEE Transactions on Circuits Systems for Video Technology, Volume 8, Issue 4, Pages 446-459, August 1998.

[26] Peng Yin, and J. Boyce, “A new rate control scheme for H.264 video coding”, Proceedings of the 2004 IEEE International Conference on Image Processing (ICIP 2004), Pages 449-452, October 2004.

[27] Garry J. Sullivan, and Thomas Wiegand, “Rate-distortion optimization for video compression”, IEEE Signal Processing Magazine, Volume 15, Issue 6, Pages 74–90, November 1998.

[28] Video Group, “Text of ISO/IEC 14496-2 MPEG-4 Video VM-Version 8.0”, ISO/IEC JTC1/SC29/WG11 Coding of Moving Pictures and Associated Audio MPEG 97/W1796, Stochholm, Sweden, July 1997.

[29] Richard Storey, Artur Alves, Jose Ruela, Luis Teixeira, and Teresa Andrade, “The ATLANTIC News Studio: Reference Model and Field Trial”, Second European Conference on Multimedia Applications, Services and Techniques (ECMAST 1997), Milan, Italy, Pages 67-78, May 1997.

[30] Nader Mohawnian, Rajesh Rajagopalan, and Cesar A. Gonzales, “Single-pass constant- and variable-bit-rate, MPEG-2 video compression”, IBM Journal of Research Development, Volume 43, Issue 4, Pages 489-509, July 1999.

[31] Sriram Sethuraman, Ravi Krishnamurthy, Xiaobing Lee, and Tihao Chiang, “Latency-based statistical multiplexing”, US Patent 6665872, December 16, 2003.

[32] Anthony Vetro, Huifang Sun, and Yao Wang, “MPEG-4 rate control for multiple video objects”, IEEE Transactions on Circuits Systems for Video Technology, Volume 9, Issue 1, Pages 186–199, February 1999.

[33] Lilla Boroczky, Agnes Y. Ngai, and Edward F. Westermann, “Statistical multiplexing using MPEG-2 video encoders”, IBM Journal of Research Development, Volume 43, Issue 4, Pages 510-520, July 1999.

[34] Jiro Katto, and Mutsumi Ohta, “Mathematical analysis of MPEG compression capability and its application to rate control”, Proceedings of the 1995 IEEE International Conference on Image Processing (ICIP 1995), Volume 2, Pages 555-558, October 1995.

[35] Limin Wang, and Andre Vincent, “Joint rate control for multi-program video coding”, IEEE Transactions on Consumer Electronics, Volume 42, Issue 3, Pages 300–305, August 1996.

[36] Limin Wang, and Andre Vincent, “Joint coding for multi-program transmission”, Proceedings of the 1996 IEEE International Conference on Image Processing (ICIP 1996), Pages 425-428, September 1996.

[37] Limin Wang, and Andre Vincent, “Bit Allocation for Joint Coding of Multiple Video Programs”, Proceedings of SPIE Volume 3024: Visual Communications and Image Processing 1997, Pages 149-158, February 1997.

[38] Limin Wang, and A. Vincent, “Bit Allocation and Constraints for Joint Coding of Multiple Video Programs”, IEEE Transactions on Circuits and Systems for Video Technology, Volume 9, Issue 6, Pages 949-959, October 1999.

[39] Lilla Boroczky, Agnes Y. Ngai, and Edward F. Westermann. “Joint rate control with look-ahead for multi–program video coding”, IEEE Transactions on Circuits Systems on Video Technology, Volume 10, Issue 7, Pages 1159–1163, October 2000.

[40] Jun Xin, Ming-Ting Sun, and KouSou Kan, “Bit allocation for joint transcoding of multiple MPEG coded video streams”, Proceedings of the 2001 IEEE International Conference on Multimedia and Expo (ICME 2001), Pages 22-25, August 2001.

[41] Zhengguo Li, Ce Zhu, Feng Pan, Gao Feng, Xiaokang Yang, S. Wu, and Nam Ling, “A Novel Joint Rate Control Scheme for the Coding of Multiple Real Time Video Programs”, Proceedings of the 22nd International Conference on Distributed Computing Systems (ICDCSW 2002), Pages 241-245, 2002.

[42] A. Vincent, P. Corriveau, P. Blanchfield, and R. Renaud, “Modeling of the Coding Gain of Joint Coding for Multi-Program Video Transmission”, Proceddings of the 2000 IEEE International Conference on Multimedia and Expo (ICME 2000), Volume 3, Pages 1309-1312, 2000.

[43] J. Jordan, and A. Bock, “Analysis, modelling and performance prediction of digital video statistical multiplexing”, Proceedings of the 1997 International Broadcasting Convention (IBC 1997), Pages 553-559, September 1997.

[44] Gertjan Keesman, and David Elias, “Analysis of joint bit-rate control in multiprogram image coding”, Proceedings of SPIE Volume 2308: Visual Communications and Image Processing 1994, Pages 1906, September 1994.

[45] Jing Yang, Xiangzhong Fang, and Hongkai Xiong, “A Joint Rate Control Scheme for H.264 Encoding of Multiple Video Sequences”, IEEE Transactions on Consumer Electronics, Volume 51, Issue 2, Pages 618-623, May 2005.

[46] Jing Yang, Xiangzhong Fang, and Hongkai Xiong, “Joint Rate Control for Multiple Sequences coding based on H.264 standard”, Proceedings of the 2005 IEEE International Conference on Multimedia and Expo (ICME 2005), Pages 133-136, July 2005.

[47] S. Sakazawa, Y. Takishima, M. Wada, and Y. Hatori, "Coding control scheme for a multiencoder system", Proceedings of the 7th International Packet Video Workshop (PV 1996), Pages 83-88, March 1996.

[48] Luís Teixeira, and Teresa Andrade, "Dynamic bandwidth allocation for an MPEG 2 multi-encoder video system", Proceedings of the Symposium on Advanced Imaging and Network Technologies - Conference on Digital Compression Technologies and Systems for Video Communications 1996, Berlin, Germany, Pages 555-566, October 1996.

[49] Luís Teixeira, "Statistical Multiplexing for TV Broadcasting", Proceedings of the 9th Portuguese Conference on Pattern Recognition (RECPAD 1997), Coimbra, Portugal, Page 23-28, March 1997.

[50] Luís Teixeira, Vítor Teixeira, and Teresa Andrade, "Dynamic Multiplexing for Digital TV Broadcasting", Proceedings of the Second European Conference on Multimedia Applications, Services and Techniques (ECMAST 1997), Milan, Italy, Pages 291-308, May 1997.

[51] Luís Teixeira, "Bit rate and buffer constraints on a joint video coding system", Proceedings of the Picture Coding Symposium 1997 (PCS 1997), Berlin, Germany, Pages 615-618, September 1997.

[52] Luís Teixeira, and Teresa Andrade, "Smoothing of MPEG Multi-program video coding for packet networks", Proceedings of the 9th International Conference on Image Analysis and Processing (ICIAP 1997), Firenze, Italy, Volume 2, Pages 117-123, September 1997.

[53] Luís Teixeira, "Global Optimisation of Jointly Video Sources", Proceedings of the 10th Portuguese Conference on Pattern Recognition (RECPAD 1998), Lisbon, Portugal, Pages 303-307, March, 1998.

[54] Luís Teixeira, "Dynamic Multiplexing of MPEG Video Streams in a Distributed Environment", Proceedings of the 6th IEEE International Workshop on Intelligent Signal Processing and Communications Systems (ISPACS 1998), Melbourne, Australia, Pages 680-684, November 1998.

[55] Mehmet Kemal Ozkan, Billy Wesley Beyers, Daniel Jorge Reininger, and Kuriacose Joseph, "Multiplexer system using constant bit rate encoders", US Patent 6055270, April 25, 2000.

[56] Limin Wang, "Rate control for MPEG video coding", Proceedings of SPIE Volume 2501: Visual Communications and Image Processing 1995, Pages 53–64, May 1995.

[57] José I. Ronda, Martina Eckert, Fernando Jaureguizar, and Narciso Garcia, "Rate control and bit allocation for MPEG-4", IEEE Transactions on Circuits Systems for Video Technology, Volume 9, Issue 8, Pages 1243–1258, December 1999.

[58] Arun N. Netravali, and Barry G. Haskell, "Digital Pictures: Representation, Compression and Standards (2nd Edition)", Plenum Press, ISBN 9780306449178, 1995.

[59] Stefan Winkler, "Digital Video Quality: Vision Models and Metrics", John Wiley & Sons, ISBN 9780470024041, 2005.

[60] Stefan Winkler, "Chapter 5: Perceptual Video Quality Metrics – A review", H. R. Wu, and K. R. Rao (Editors), "Digital Video Image Quality and Perceptual Coding", Eds. Boca Raton, FL: CRC Press, ISBN 0824727770, November 2005.

[61] Joana Sofia Cardoso Palhais, "Quality of Experience Assessment in Internet TV", Master Dissertation, Instituto Superior Tecnico, May 2011.

[62] Marcio N. Zapater, and Graça Bressan, "A Proposed Approach for Quality of Experience Assurance of IPTV", Proceedings of the First International Conference on the Digital Society (ICDS'07), Page 25, January 2007.

[63] ITU-T Recommendation P.10/G.100, “Amendment 2: New definitions for inclusion in Recommendation ITU-T P.10/G.100”, Geneva, Switzerland, July 2008.

[64] Kalpana Seshadrinathan, Thrasyvoulos N. Pappas, Robert J. Safranek, Junqing Chen, Zhou Wang, Hamid R. Sheikh, and Alan C. Bovik, “Image quality assessment”, Alan C. Bovik (Editor), “Essential Guide to Image Processing”, Elsevier, ISBN 9780123744579, 2009.

[65] ITU-T Recommendation E.800, “Definitions of terms related to quality of service”, Geneva, Switzerland, September 2008.

[66] Junyong You, Ulrich Reiter, Miska M. Hannuksela, Moncef Gabbouj, and Andrew Perkis, “Perceptual-based quality assessment for audio-visual services: A survey”, Journal of Image Communication, Volume 25, Issue 7, Pages 482-501, August 2010.

[67] Kalpana Seshadrinathan, and Alan C. Bovik, “New vistas in image and video quality assessment”, Proceedings of SPIE Volume 6492: Human Vision and Electronic Imaging XII, February 2007.

[68] Hamid Rahim Sheikh, and Alan C. Bovik, “Image Information and Visual Quality”, IEEE Transactions on Image Processing, Volume 15, Issue 2, Pages 430-444, February 2006.

[69] Stefan Winkler, and Praveen Mohandas, "The Evolution of Video Quality Measurement: From PSNR to Hybrid Metrics”, IEEE Transactions on Broadcasting, Volume 54, Issue 3, Pages 660-668, September 2008.

[70] ITU-R Recommendation BT.500-7, “Methodology for the subjective assessment of the quality of television pictures”, Geneva, Switzerland, 1995.

[71] ITU-R Recommendation BT.500-12, “Methodology for the subjective assessment of the quality of television pictures”, Geneva, Switzerland, 2009.

[72] ITU-T Recommendation P.r910, “Subjective video quality assessment methods for multimedia applications”, Geneva, Switzerland, 2008.

[73] ITU-T Recommendation P.911, “Subjective Audiovisual Quality Assessment Methods for Multimedia Applications”, Geneva, Switzerland, 1998.

[74] Roger N. Shepard, A. Kimball Romney, and Sara Beth Nerlove (Editors), “Multidimensional Scaling: Theory and Applications in the Behavioral Sciences”, New York Seminar Press, ISBN 0129104027, 1972.

[75] D. E. Pearson, “A Three-Stage Process for the Evaluating of Image Quality”, Proceedings of the Society for Information Display, Volume 21, Issue 3, Pages 271-278, 1980.

[76] D. E. Pearson, “Methods for Scaling Television Picture quality: A Survey”, Thomas S. Huang and Ohl J. Tretiak (Editors), Symposium on Picture Bandwidth Compression, Gordon and Breach, New York, Pages 47-95, 1975.

[77] Stefan Winkler, “Video quality measurement standards - current status and trends”, Proceedings of the 7th International Conference on Information, Communications and Signal Processing (ICICS 2009), Macau, Pages 1-5, December 2009.

[78] VQEG Full Reference Television (FRTV) Phase I Group, “Final report from the Video Quality Experts Group on the validation of objective models of video quality assessment”, March 2000. Available from: http://www.its.bldrdoc.gov/vqeg/projects/frtv-phase-i/frtv-phase-i.aspx. Retrieved 20 January 2012.

[79] Kjell Brunnström, David Hands, Filippo Speranza, and Arthur Webster, “Standards in a Nutshell - VQEG Validation and ITU Standardisation of Objective Perceptual Video Quality Metrics”, IEEE Signal Processing Magazine, Volume 26, Issue 3, Pages 96-99, May 2009.

[80] Matthew D. Brotherton, Quan Huynh-Thu, David S. Hands, and Kjell Brunnström, “Subjective multimedia quality assessment”, IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, Volume E89-A, Issue 11, Pages 2920–2932, November 2006.

[81] SAMVIQ - Subjective Assessment Methodology for VIdeo Quality, EBU Project Group B/VIM Video In Multimedia, May 2003.

[82] Franc Kozamernik, Paola Sunna, Emmanuel Wyckens, and Dag Inge Pettersen, “Subjective quality of internet video codecs — phase II evaluations using SAMVIQ”, EBU Technical Review, Number 301, January 2005.

[83] ITU-R Report BT.1082-1, “Studies toward the unification of picture assessment methodology”, Geneva, Switzerland, January 1990.

[84] ITU-T Recommendation G.1011, “Reference guide to Quality of Experience (QoE) assessment methodologies”, 2010, based on ITU-T G.RQAM, “Reference guide to QoE assessment methodologies”, Standard Draft TD 310rev1, May 2010.

[85] Shyamprasad Chikkerur, Vijay Sundaram, Martin Reisslein, and Lina J. Karam, “Objective Video Quality Assessment Methods: A Classification, Review, and Performance Comparison”, IEEE Transactions on Broadcasting, Volume 57, Issue 2, Pages 165-181, June 2011.

[86] Sebastian Möller, and Alexander Raake, “Telephone Speech Quality Prediction: Towards Network Planning and Monitoring Models”, Speech Communication, Volume 38, Issue 1, Pages 47-75, September 2002.

[87] Akira Takahashi, David Hands, and Vincent Barriac, “Standardisation Activities in the ITU for a QoE Assessment of IPTV”, IEEE Communications Magazine, Volume 46, Issue 2, Pages 78-84, February 2008.

[88] David Hands, “Quality Assurance for IPTV”, ITU-T Workshop.End-to-End QoE/QoS, Geneva, Switzerland, June 2006.

[89] Ilona Roth, “An introduction to object perception”, Ilona Roth, and John P. Frisby (Editors), “Perception and representation. A Cognitive Approach”, Milton Keynes, ISBN 0335153283, Pages 81-106, 1986.

[90] Ian Stuart-Hamilton, “Key ideas in Psychology”, Jessica Kingsley Publishers, ISBN 9781853023590, 1999.

[91] E. Bruce Goldstein, “Sensation and Perception (7th Edition)”, Thomson, ISBN 049518778X, 2007.

[92] Douglas W. Bloomquist, “Teaching sensation and perception: Its ambiguous and subliminal aspects”, Anne M. Rogers, and C.J. Scheirer (Editors), The G. Stanley Hall Lecture Series, American Psychological Association, Washington, DC, Volume 5, Pages 159-203, 1985.

[93] Ian E. Gordon, “Theories of Visual Perception (3rd Edition)”, Psychology Press, ISBN 1841693839, September 2004.

[94] Victoria Johnstone, and Brent Alsop, “Human signal-detection performance: Effects of signal presentation probabilities and reinforcer distributions”, Journal of the Experimental Analysis of Behavior, Volume 66, Pages 243–263, 1996.

[95] Brian A. Wandell, “Foundations of Vision”, Sinauer Associates, Inc, ISBN 0878938532, May 1995.

[96] Gordon E. Legge, and John M. Foley, “Contrast masking in human vision”, Journal of the Optical Society of America, Volume 70, Issue 12, Pages 1458–1471, 1980.

[97] Jeffrey Lubin, “A visual discrimination model for image system design and evaluation”, E. Peli (Editor), “Visual Models for Target Detection and Recognition”, World Scientific Publisher, Singapore, Pages 207-220, 1995.

[98] Patrick C. Teo, and David J. Heeger, “Perceptual image distortion”, Proceedings of the 1994 IEEE International Conference on Image Processing (ICIP 1994), Volume 2, Pages 982-986, 1994.

[99] John M. Foley, “Human luminance pattern-vision mechanisms: masking experiments require a new model”, Journal of Optical Society of America A, Volume 11, Issue 6, Pages 1710-1719, 1994.

[100] M. A. Losada, and K. T. Mullen, “The spatial tuning of chromatic mechanisms identified by simultaneous masking”, Vision Research, Volume 34, Issue 3, Pages 331-341, February 1994.

[101] Eugene Switkes, Arthur Bradle, and Karen K. De Valois, “Contrast dependence and mechanisms of masking interactions among chromatic and luminance gratings”, Journal Optical Society of America A, Volume 5, Issue 7, Pages 1149-1162, 1988.

[102] Wa James Tam, Lew B. Stelmach, Limin Wang, Daniel Lauzon, and Peter Gray, “Visual masking at video scene cuts”, Proceedings of SPIE Volume 2411: Human Vision Visual Processing and Digital Display VI, Pages 111-119, February 1995.

[103] Albert J. Ahumada Jr., Bettina L. Beard, and Robert Eriksson, “Spatio-temporal discrimination model predicts temporal masking function”, Proceedings of SPIE Volume 3299: Human Vision and Electronic Imaging III, Pages 120–127, 1998.

[104] Bernd Girod, “The information theoretical significance of spatial and temporal masking in video signals”, Proceedings of SPIE Volume 1077: Human Vision, Visual Processing and Digital Display, Pages 178–187, 1989.

[105] R. E. Fredericksen, and R. F. Hess, “Estimating multiple temporal mechanisms in human vision”, Vision Research, Volume 38, Issue 7, Pages 1023-1040, 1998.

[106] R. F. Quick Jr., “A vector-magnitude model of contrast detection”, Kybernetik, Volume 16, Pages 65-67, 1974.

[107] Huib De Ridder, “Minkowski-metrics as a combination rule for digital-image-coding impairments”, Proceedings of SPIE Volume 1666: Vision, Visual Processing and Digital Display III, Pages 16-26, 1992.

[108] Jeffrey Lubin, “The use of psychophysical data and models in the analysis of display system performance”, Andrew B. Watson (Editor), “Digital Images and Human Vision”, The MIT Press, ISBN 0262231719, Pages 163-178, 1993.

[109] Andrew B. Watson, “DCT quantisation matrices visually optimized for individual images”, Proceedings of SPIE Volume 1913: Human Vision, Visual Processing and Digital Display IV, Pages 202–216, February 1993.

[110] Christian J. van den Branden Lambrech, and Olivier Verscheure, “Perceptual quality measure using a spatio-temporal model of the human visual system”, Proceedings of the SPIE Volume 2668: Digital Video Compression: Algorithms and Technologies 1996, Pages 450-461, 1996.

[111] Wen Xu, and Gert Hauske, “Picture quality evaluation based on error segmentation”, Proceedings of SPIE Volume 2308: Visual Communications and Image Processing 1994, Pages 1454-1465, 1994.

[112] Wilfried Osberger, Neil Bergmann, and Anthony Maeder, “An automatic image quality assessment technique incorporating high level perceptual factors”, Proceedings of the 1998 IEEE International Conference Image Processing (ICIP 1998), Volume 3, Pages 414-418, 1998.

[113] Thrasyvoulos N. Pappas, Thomas A. Michel, and Raynard O. Hinds, “Supra-threshold perceptual image coding”, Proceedings of the 1996 IEEE International Conference on Image Processing (ICIP 1996), Lausanne, Switzerland, Volume I, Pages 237-240, September 1996.

[114] R. W. G. Hunt, “The reproduction of colour (6th Edition)”, Wiley, ISBN 0470024259, 2004.

[115] Marcus Nadenau, “Integration of human color vision models into high quality image compression”, Ph.D Dissertation, Signal Processing Laboratory, Swiss Federal Institute of Technology, Lausanne, November 2000.

[116] Thrasyvoulos N. Pappas, and David L. Neuhoff, “Least-squares model-based halftoning”, IEEE Transaction on Image Processing, Volume 8, Issue 8, Pages 1102-1116, August 1999.

[117] James L. Mannos, and David J. Sakrison, “The Effects of a Visual Fidelity Criterion on the Encoding of Images”, IEEE Transactions on Information Theory, Volume 20, Issue 4, Pages 525-536, July 1974.

[118] F. W. Campbell, J. J. Kulikowski, and J. Levinson, “The effect of orientation on the visual resolution of gratings”, Journal of Physiology, Volume 187, Issue 2, Pages 427–436, November 1966.

[119] L. A. Olzak, and J. P. Thomas, “Chapter 7: Seeing spatial patterns”, Kenneth R. Boff, Lloyd Kaufman, James P. Thomas (Editors), “Handbook of perception and human performance”, Wiley-Interscience, New York, ISBN 0471885444, Volume 1, Pages 1–55, May 1986.

[120] Stanley A. Klein, Amnon D. Silverstein, and Thom Carney, “Relevance of human vision to JPEG-DCT compression”, Proceedings of 1992 SPIE Volume 1666: Human Vision, Visual Processing and Digital Display III, Pages 200–213, 1992.

[121] Benoit Macq, “Weighted optimum bit allocations to orthogonal transforms for picture coding”, IEEE Journal on Selected Areas in Communications, Volume 10, Issue 5, Pages 875–883, June 1992.

[122] Norman B. Nill, “A Visual Model Weighted Cosine Transform for Image Compression and Quality Assessment”, IEEE Transactions on Communications, Volume 33, Issue 6, Pages 551–557, June 1985.

[123] Heidi A. Peterson, Albert J. Ahumada Jr., and Andrew B. Watson, “An improved detection model for DCT coefficient quantisation”, Proceedings of SPIE Volume 1913: Human Vision, Visual Processing and Digital Display IV, Pages 191–201, February 1993.

[124] Yun-Chin Li, Tong-Hai Wu, and Yung-Chang Chen, “A scene adaptive hybrid video coding scheme based on the LOT”, IEEE Transactions on Circuits and Systems for Video Technology, Volume 8, Issue 1, Pages 92–103, February 1998.

[125] Robert J. Safranek, “JPEG compliant encoder utilizing perceptually based quantisation”, Proceedings of SPIE Volume 2179: Human Vision, Visual Processing and Digital Display V, Pages 117–126, May 1994.

[126] Andrew B. Watson, James Hu, and John F. McGowan III, “DVQ: A digital video quality metric based on human vision”, Journal of Electronic Imaging, Volume 10, Issue 1, Pages 20-29, 2001.

[127] Christian J. van den Branden Lambrecht, “Perceptual models and architectures for video coding applications”, Ph.D Dissertation, Swiss Federal Institute of Technology, August 1996.

[128] Nikil Jayant, James Johnston, and Robert Safranek, “Signal compression based on models of human perception”, Proceedings of the IEEE, Volume 81, Pages 1385-1422, October 1993.

[129] Andrew B. Watson, “The cortex transform: rapid computation of simulated neural images”, Computer Vision, Graphics, and Image Processing, Volume 39, Pages 311-327, 1987.

[130] Scott J. Daly, “The visible difference predictor: An algorithm for the assessment of image fidelity”, Proceedings of SPIE Volume 1616: Human Vision, Visual Processing, and Digital Display III, Pages 2-15, August 1992.

[131] Patrick C. Teo, and David J. Heeger, “Perceptual image distortion”, Proceedings of SPIE Volume 2179: Human Vision, Visual Processing and Digital Display IV, Pages 127–141, February 1994.

[132] David J. Heeger, and Patrick C. Teo, “A model of perceptual image fidelity”, Proceedings of the the 1995 IEEE International Conference in Image Processing (ICIP 1995), Pages 343-345, 1995.

[133] Peter J. Burt, and Edward H. Adelson, "The Laplacian pyramid as a compact image code", IEEE Transactions Communication, Volume 31, Issue 4, Pages 532-540, 1983.

[134] William T. Freeman, and Edward H. Adelson, "The design and use of steerable filters", IEEE Transactions Pattern Analysis and Machine Intelligent, Volume 13, Issue 9, Pages 891-906, September 1991.

[135] Eero P. Simoncelli, William T. Freeman, Edward H. Adelson, and David J. Heeger, "Shiftable multiscale transforms", IEEE Transactions on Information Theory, Volume 38, Issue 2, Part 2, Pages 587-607, March 1992.

[136] Andrew B. Watson, and Joshua A. Solomon, "Model of visual contrast gain control and pattern masking", Journal of Optical Society of America. A, Volume 14, Issue 9, Pages 2379-2391, 1997.

[137] F. W. Campbell, and R. W. Gubisch, "Optical quality of the human eye", Journal of Physiology, Volume 186, Issue 3, Pages 558–578, March 1966.

[138] D. H. Kelly, "Adaptation effects on spatio-temporal sinewave thresholds", Vision Research, Volume 12, Issue 1, Pages 89–101, January 1972.

[139] John G. Nicholls, A. Robert Martin, and Bruce G. Wallace, "From neuron to brain: a cellular approach to the functions of the nervous system", Sunderland, Mass. Sinauer Associates, ISBN 087893586X, 1992.

[140] Junqing Chen, and N. Thrasyvoulos Pappas, "Perceptual Metrics and Perceptual Coders", Proceedings of SPIE Volume 4299: Human Vision and Electronic Imaging VI, Pages 150-162, January 2001.

[141] Jacob Nachmias, and Richard Sansbury, "Grating contrast: discrimination may be better than detection", Vision Research, Volume 14, Issue 10, Pages 1039-1042, October 1974.

[142] G. C. DeAngelis, J. G. Robson, I. Ohzawa, and R. D. Freeman, "Organization of suppression in receptive fields of neurons in cat visual cortex", Journal of Neurophysiology, Volume 68, Issue 1, Pages 144-163, July 1992.

[143] Wilson S. Geisler, and Duane G. Albrecht, "Cortical neurons: Isolation of contrast gain control", Vision Research, Volume 32, Issue 8, Pages 1409-1410, August 1992.

[144] David J. Heeger, "Normalization of cell responses in cat striate cortex", Visual Neuroscience, Volume 9, Issue 2, Pages 181-197, 1992.

[145] Hugh R. Wilson and Richard Humanski, "Spatial frequency adaptation and contrast gain control", Vision Research, Volume 33, Issue 8, Pages 1133-1149, May 1993.

[146] Ralph E. Jacobson, "An evaluation of image quality metrics", Journal of Photographic Science, Volume 43, Issue 1, Pages 7-16, 1995.

[147] Albert J. Ahumada, and Cynthia E. Null, "Image Quality: a multidimensional problem", A. B. Watson (Editor), Digital Images and Human Vision, Cambridge MA: MIT Press, Pages 141-148, 1993.

[148] G. P. Corey, M. J. Clayton, and K. N. Cupery, "Scene dependence of image quality", Society of Photographic Scientists and Engineers, Volume 27, Pages 9-13, January-February 1983.

[149] Gregory R. Lockhead, "Psychophysical scaling: judgements of attributes or objects?", Behavioral and Brain Sciences Volume 15, Issue 3, Pages 543-601, September 1992.

[150] Rafael C. Gonzalez, and Richard E. Woods, "Digital Image Processing (3rd Edition)", Prentice Hall, ISBN 013168728X, August 2007.

[151] Anil K. Jain, "Fundamentals of Digital Image Processing", Prentice Hall, ISBN 0133361659, 1989.

[152] Daniel R. Fuhrman, John A. Baro, and Jerome R. Cox, “Experimental evaluation of psychophysical distortion metrics for JPEG-encoded images”, J. P. Allebach, and R. E. Rogowitz (Editors), Proceedings of SPIE Volume 1913: Human Vision, Visual Processing, and Digital Display IV, Pages 179-190, 1993.

[153] Arthur A. Webster, Coleen T. Jones, Margaret H. Pinson, Stephen D. Voran, and Stephen Wolf, “An objective video quality assessment system based on human perception”, J. P. Allebach, and R. E. Rogowitz (Editors), Proceedings of SPIE Volume 1913: Human Vision, Visual Processing, and Digital Display IV, Pages 15-26, 1993.

[154] Bernd Girod, “What’s wrong with mean-squared error?”, Andrew B. Watson (Editor), “Digital Images and Human Vision”, MIT Press, ISBN 026223171, Pages 207-220, 1993.

[155] Stanley A. Klein, “Image quality and image compression: a psychophysicist’s viewpoint”, Andrew B. Watson (Editor), “Digital Images and Human Vision”, MIT Press, ISBN 026223171, Pages 73-88, 1993.

[156] Stefan Winkler, “Video quality and beyond”, Proceedings of the 2007 European Signal Processing Conference (EUSIPCO 2007), Poznañ, Poland, September 2007.

[157] Jeffrey Lubin, and David Fibush, “Sarnoff JND Vision Model”, T1A1.5 Working Group Document #97-612, ANSI T1 Standards Committee, 1997.

[158] ANSI T1.801.03-1996, “American National Standards for Telecommunications - Digital Transport of One-Way Digital Signals – Parameters for Objective Performance Assessment”, Alliance for Telecommunications Industry Solutions, 1200 G Street, NW, Suite 500,Washington DC 20005, 1996.

[159] ITU-T Recommendation J.143, “User requirements for objective perceptual video quality measurements in digital cable television”, International Telecommunications Union, Geneva, Switzerland, 2000.

[160] Tomas Brandão, and Maria P. Queluz, "No-reference perceptual quality metric for H.264/AVC encoded video", Proceedings of the Fitth International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM 2010), Scottsdale, United States, January, 2010.

[161] Stefan Winkler, Animesh Sharma, and David McNally, “Perceptual video quality and blockiness metrics for multimedia streaming applications”, Proceedings of the Fourth International Symposium on Wireless Personal Multimedia Communication (WPMC 2001), Aalborg, Denmark, Pages 547–552, September 2001.

[162] Seyed Ali Amirshahi, “Towards a Perceptual Metric for Video Quality Assessment”, Master Thesis, Gjovik University College, Norway, June 2010.

[163] Zhou Wang, and Alan C. Bovik, “Modern Image Quality Assessment”, Synthesis Lectures on Image, Video, and Multimedia Processing, Morgan and Claypool Publishing Company, Volume 2, Number 1, Pages 1-156, 2006.

[164] Zhou Wang, Alan C. Bovik, and Eero P. Simancelli, “Chapter 8.3: Structural approaches to image quality assessment”, A. Bovik (Editor), Handbook of Image and Video Processing (2nd Edition), Elsevier Academic Press, Amsterdam, The Netherlands, 2005.

[165] ANSI T1.801.01-1995, “American National Standards for Telecommunications - Digital Transport of Video Teleconferencing/Video Telephony Signals - Video Test Scenes for Subjective and Objective Performance Assessment”, Alliance for Telecommunications Industry Solutions, 1200 G Street, NW, Suite 500,Washington DC 20005, 1995.

[166] ANSI T1.801.02-1996, “American National Standards for Telecommunications - Digital Transport of Video Teleconferencing/Video Telephony Signals - Performance Terms, Definitions and Examples”, Alliance for Telecommunications Industry Solutions, 1200 G Street, NW, Suite 500,Washington DC 20005, 1996.

[167] David Fibush, “Overview of Picture Quality Measurement Methods”, IEEE G-2.1.6 Compression and Processing Subcommittee, May 12, 1997.

[168] Technical Subcommittee T1A1.5, “ Digital Transport of One-Way Video Signals - Parameters for Objective Performance Assessment”, T1.801.03-1996, ANSI, February 1996.

[169] Keng-Pang Lim, Garry Sullivan, and Thomas Wiegand, “Text Description of Joint Model Reference Encoding Methods and Decoding Concealment Methods”, Joint Video Team of ISO/IEC MPEG and ITU-T VCEG, Document JVT-W057, San Jose, USA, April 2007.

[170] H.264/AVC JM Reference Software. Available from: http://iphome.hhi.de/suehring/tml/. Retrieved 31 January 2012.

[171] John M. Libert, Leon Stanger, Andrew B. Watson, and Ann M. Rohaly, “Toward developing a unit of measure and scale of digital video quality: IEEE Broadcast Technology Society Subcommittee on Video Compression Measurements”, Proceedings of SPIE Volume 3959: The International Society for Optical Engineering, Pages 160-165, January 2000.

[172] ATIS Technical Report T1.TR.72-2001, “Methodological framework for specifying accuracy and cross-calibration of video quality metrics”, Alliance for Telecommunications Industry Solutions, 1200 G Street, NW, Suite 500, Washington, DC 20005, October 2001.

[173] ATIS Technical Report T1.TR.73-2001, “Video normalization methods applicable to objective video quality metrics utilizing a full reference technique”, October 2001.

[174] ATIS Technical Report T1.TR.74-2001, “Objective video quality measurement using a peak-signal-to-noise-ratio (PSNR) full reference technique”, October 2001.

[175] ATIS Technical Report T1.TR.75-2001, “Objective video quality measurement using a JND-based full reference technique”, October 2001.

[176] VQEG Full Reference Television (FRTV) Phase II Group, “Final report from the Video Quality Experts Group on the validation of objective models of video quality assessment, Phase II”, August 2003. Available from: http://www.its.bldrdoc.gov/vqeg/projects/frtv-phase-ii/frtv-phase-ii.aspx. Retrieved 31 January 2012.

[177] VQEG Full Reference Television (FRTV) Phase II Group, “Full Reference Television, Phase II, Subjective Test Plan (version 1.7)”, September 2002. Available from: http://www.its.bldrdoc.gov/vqeg/projects/frtv-phase-ii/frtv-phase-ii.aspx. Retrieved 31 January 2012.

[178] ITU-R Recommendation BT.1683, “Objective perceptual video quality measurement techniques for standard definition digital broadcast television in the presence of a full reference”, Geneva, Switzerland, 2004.

[179] VQEG Multimedia (MM) Phase I Group, “Final report from the Video Quality Experts Group on the validation of objective models of multimedia quality assessment, Phase I (Version 2.6)”, September 2008. Available from: http://www.its.bldrdoc.gov/vqeg/projects/multimedia-phase-i/multimedia-phase-i.aspx. Retrieved 31 January 2012.

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

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

[182] VQEG Reduced-Reference and No-Reference (RRNR-TV) Phase I Group, “Final report from the Video Quality Experts Group on the validation of Reduced-Reference and No-Reference Objective Models for

Standard Definition Television, Phase I (Version 1.8)", 2009. Available from: http://www.its.bldrdoc.gov/vqeg/projects/rrnr-tv/rrnr-tv.aspx. Retrieved 31 January 2012.

[183] VQEG High Definition Television (HDTV) Phase I Group, "Report on the Validation of Video Quality Models for High Definition Video Content (Version 2.0)", June 2010, Available from: http://www.its.bldrdoc.gov/vqeg/projects/hdtv/hdtv.aspx. Retrieved 31 January 2012.

[184] Nikil Jayant, "Signal compression: Technology targets and research directions", IEEE Journal in Selected Areas in Communications, Volume 10, Issue 5, Pages 796-818, June 1992.

[185] Andrew B. Watson, Gloria Y. Yang, Joshua A. Solomon, and John Villasenor, "Visibility of wavelet quantisation noise", IEEE Transactions on Image Processing, Volume 6, Issue 8, Pages 1164–1175, August 1997.

[186] Ingo Hontsch, and Lina J. Karam, "Adaptive image coding with perceptual distortion control", IEEE Transactions on Image Processing, Volume 11, Issue 3, Pages 213–222, March 2002.

[187] Chun-Hsien Chou, and Yun-Chin Li, "A perceptually tuned subband image coder based on the measure of just-noticeable-distortion profile", IEEE Transactions on Circuits Systems for Video Technology, Volume 5, Issue 6, Pages 467–476, June 1995.

[188] Chun-Hsien Chou, and Chi-Wei Chen, "A perceptually optimized 3-D subband image codec for video communication over wireless channels", IEEE Transactions on Circuits Systems for Video Technology, Volume 6, Issue 2, Pages 143–156, February 1996.

[189] Chi-Chang Kuoa, and Jin-Jang Leou, "A New Rate Control Scheme for H.263 Video Transmission", Signal Processing: Image Communication, Volume 17, Issue 7, Pages 537-557, August 2002.

[190] Ishfaq Ahmad, and Jiancong Luo, "On Using Game Theory to Optimize the Rate Control in Video Coding", IEEE Transaction on Circuits and Systems for Video Technology, Volume 16, Issue 2, Pages 209-219, February 2006.

[191] X. K. Yang, W. S. Lin, Zhongkang Lu, E. P. Ong, and Susu S. Yao, "Just-noticeable-distortion profile with nonlinear additivity model for perceptual masking in color images", Proceedings of the 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2003), Hong Kong, Volume 3, Pages 609–612, April 2003.

[192] John Canny, "A computational approach to edge detection", IEEE Transactions on Pattern Analysis and Machine Intelligent, Volume 8, Issue 6, Pages 679–698, 1986.

[193] Wang Zhou, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli, "Image Quality Assessment: From Error Measurement to Structural Similarity", IEEE Transactions on Image Processing, Volume 13, Issue 4, Pages 600-613, April 2004.

[194] Alan C. Brooks, and Thrasyvoulos N. Pappas, "Structural similarity quality metrics in a coding context: exploring the space of realistic distortions", IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2007), Honolulu, Hawaii, Pages 869-872, April 2007.

[195] Anush K. Moorthy, Zhou. Wang, and Alan C. Bovik, "Chapter 19: Visual perception and quality assessment", Gabriel Cristobal, Peter Schelkens, and Hugo Thienpont (Editors), "Optical and Digital Image Processing: Fundamentals and Applications", John Wiley & Sons, ISBN 9783527409563, 2011.

[196] Guan-Lin Wu, Tung-Hsing Wu, and Shao-Yi Chien, "Algorithm and Architecture Design of Perception Engine for Video Coding Applications", IEEE Transactions on Multimedia, Volume 13, Issue 6, Pages 1181-1194, December 2011.

[197] Zhou Wang, and Alan C. Bovik, "Embedded foveation image coding", IEEE Transactions on Image Processing, Volume 10, Issue 10, Pages 1397-1410, October 2001.

[198] Anush K. Moorthy, and Alan C. Bovik, "Perceptually significant spatial pooling techniques for image quality assessment", Proceedings of SPIE Volume 7240: Human Vision and Electronic Imaging XIV, Pages 724012-1, 2009.

[199] Anush K. Moorthy, and Alan C. Bovik, "Visual importance pooling for image quality assessment", IEEE Journal of Selected Topics in Signal Processing, Volume 3, Issue 2, Issue on Visual Media Quality Assessment, Pages 193–201, 2009.

[200] Zhou Wang, and Xinli Shang, "Spatial pooling strategies for perceptual image quality assessment", IEEE International Conference on Image Processing 2006 (ICIP 2006), Atlanta, U.S.A., Pages 2945-2948, September 2006.

[201] Chaofeng Li, and Alan C. Bovik, "Three-Component Weighted Structural Similarity Index", Proceedings of SPIE Volume 7242: Image Quality and System Performance VI, Pages 72420Q, 2009.

[202] Thomas Sporer, Jens-Oliver Fischer, Judith Liebetrau, Daniel Fröhlich, Sebastian Schneider, and Sven Kämpf, "D3 Study: Final report - Definition of an objective evaluation method for assessing the minimal quality of digital video and audio sources required to provide simultaneous interpretation", version 1.7, December 2010.

[203] Zhou Wang, and Eero P. Simoncelli, "Translation insensitive image similarity in complex wavelet domain", Proceedings of the 2005 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2005), Volume 2, Pages 573-576, 2005.

[204] Mehul P. Sampat, Zhou Wang, Shalini Gupta, Alan Conrad Bovik, and Mia K. Markey, "Complex wavelet structural similarity: a new image similarity index", IEEE Transactions on Image Processing, Volume 18, Issue 11, Pages 2385-2401, November 2009.

[205] Shalini Gupta, Mia K Markey, and Alan C Bovik, "Advances and challenges in 3D and 2D+3D human face recognition", Marsha S. Corrigan (Editor), "Pattern Recognition in Biology", Nova Publishers, ISBN 1600217168, Pages 63-103, 2007.

[206] Zhou Wang, Eero P. Simoncelli, Alan C. Bovik, "Multi-scale structural similarity for image quality assessment", Proceedings of the 37th IEEE Asilomar Conference on Signals, Systems and Computers, Volume 2, Pages 1398-1402, November 2003.

[207] Weisi Lin, and C.-C. Jay Kuo, "Perceptual visual quality metrics: A survey", Journal of Visual Communication and Image Representation, Volume 22, Issue 4, Pages 297-312, May 2011.

[208] Guan-Hao Chen, Chun-Ling Yang, and Sheng-Li Xie, "Gradient-based structural similarity for image quality assessment", IEEE International Conference on Image Processing (ICIP 2006), Pages 2929–2932, 2006.

[209] Xinbo Gao, Tao Wang, and Jie Li, "A content-based image quality metric", Lecture Notes in Computer Science, Volume 3642, Pages 231-240, 2005.

[210] Srivatsan Kandadai, Joseph Hardin, and Charles D. Creusere, "Audio quality assessment using the mean structural similarity measure", Proceedings of the 2008 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2008), Las Vegas, U.S.A., Pages 221–224, March–April 2008.

[211] Margaret H. Pinson, and Stephan Wolf, "A new standardized method for objectively measuring video quality", IEEE Transactions Broadcasting, Volume 50, Issue 3, Pages 312–322, September 2004.

[212] David S. Hands, Damien Bayart, Andrew Davis, and Alex Bourret, “No Reference Perceptual Quality Metrics: Approaches and limitations”, Proceedings of SPIE Volume 7240: Human Vision and Electronic Imaging XIV, Pages 72400Y, January 2009.

[213] CCITT Recommendation H.120, “Codecs for videoconferencing using primary digital group transmission”, Geneva, Switzerland, 1989.

[214] ITU-T Recommendation H.261, “Video codec for Audiovisual Services at p x 64 kbit/s”, Geneva, Switzerland, March 1993.

[215] ITU-T Recommendation H.263, “Video Coding for Low bit rate Communication”, version 1, November 1995; version 2, January 1998, version 3, November 2000.

[216] ISO/IEC 10918-1:1994, “Information Technology - Digital compression and coding of continuous-tone still images: requirements and guidelines”, 1994.

[217] MPEG, “The MPEG vision”, ISO/IEC JTC1/SC29/WG11, Document N10412, Lausanne MPEG meeting, February 2009.

[218] Didier Le Gall, “MPEG: A Video Compression Standard for Multimedia Applications”, Communications of the ACM, Volume 34, Issue 4, Pages 47-58, April 1991.

[219] Sakae Okubo, “Reference Model Methodology - A Tool for the Collaborative Creation of Video Coding Standards”, Proceedings of the IEEE, Volume 83, Issue 2, Pages 139-150, February 1995.

[220] Ralf Schafer, and Thomas Sikora, “Digital Video Coding Standards and their Role in Video Communications”, Proceedings of the IEEE, Volume 83, Issue 6, Pages 907-924, June 1995.

[221] ISO/IEC 11172-2:1993, "Information Technology-Coding of Moving Pictures and Associated Audio for Digital Storage Media at up to about 1.5 Mbit/s - Part 2: Video", 1993.

[222] ISO/IEC JTC1/SC29/WG11, “MPEG-4 Additional Call for Proposals and Proposal Package Description”, Document N1108, Dallas MPEG meeting, November 1995.

[223] JVT Document Archive Site. Available from: http://ftp3.itu.ch/av-arch/jvt-site. Retrieved 31 January 2012.

[224] José M. Martínez, “MPEG-7 Overview (version 10)”, ISO/IECJTC1/SC29/WG11, Document N6828, Palma de Mallorca MPEG meeting, October 2004.

[225] MPEG Requirements Group, “MPEG-21 Overview v.5”, ISO/IEC JTC1/SC29/WG11, Document N5231, Shanghai MPEG meeting, October 2002.

[226] ISO/IEC TR 21000-1:2001(E) Part 1: Vision, Technologies and Strategy. Available from: http://www.iso.ch/iso/en/ittf/PubliclyAvailableStandards. Retrieved 31 January 2012.

[227] ISO/IEC JTC1, “Coding of audio-visual objects – Part 2: Visual”, ISO/IEC 14496-2 (MPEG-4 visual version 1), April 1999; Amendment 1 (version 2), February, 2000; Amendment 4 (streaming profile), January 2001.

[228] Phil N. Tudor, “MPEG-2 Video compression tutorial”, IEE Colloquium on MPEG-2 - What it is and What it isn't, Pages 2/1 – 2/8, January 1995.

[229] MPEG-2 Video Requirements Subgroup, “Agreements on Profile/Level”, ISO/IEC JTC1/SC29/WG11, Document N0489, New York MPEG meeting, July 1993.

[230] Rob Koenen, and Fernando Pereira, “MPEG-4: a new way towards multimedia mobile communications”, RACE Mobile Summit, Cascais-Portugal, November 1995.

[231] ITU-T Recommendation H.263++, “Video Coding for Low Bit Rate Communications”, Geneva, Switzerland, 2000.

[232] Thomas Sikora, “The MPEG-4 Video Standard Verification Model”, IEEE Transactions on Circuits and Systems for Video Technology, Volume 7, Issue 1, Pages 19-31, February 1997.

[233] Fernando Pereira, and Thierry Alpert, “MPEG-4 Video Subjective Tests Procedures and Results”, IEEE Transactions on Circuits and Systems for Video Technology, Volume 7, Issue 1, Pages 32-51, February 1997.

[234] MPEG Applications and Operational Environments Group, “Proposal Package Description (PPD): Revision 3”, ISO/IEC JTC1/SC29/WG11, Document N998, Tokyo MPEG meeting, July 1995.

[235] ISO/IEC JTC 1/SC 29, “Programme of Work – MPEG-4 (Coding of audio-visual objects)”, November 2009.

[236] ISO/IEC 14496-1:2010, “Information Technology – Coding of Audio-Visual Objects – Part 1: Systems (3rd Ed.)”, 2010.

[237] ISO/IEC 14496-1:2010/Amd 1:2010, “Usage of LASeR in MPEG-4 systems and Registration Authority for MPEG-4 descriptors (1st Ed.)”, 2010.

[238] ISO/IEC 14496-2:2004, “Information Technology – Coding of Audio-Visual Objects – Part 2: Visual (3rd Ed.)”, 2004

[239] ISO/IEC 14496-2:2004/Amd 5:2009, “Simple studio profile levels 5 and 6 (1st Ed.)”, 2009.

[240] ISO/IEC 14496-3:2009, “Information technology — Coding of audio-visual objects — Part 3: Audio (4th Ed.)”, 2009.

[241] ISO/IEC 14496-3:2009/Amd 2:2010, “ALS simple profile and transport of SAOC (1st Ed.)”. 2010.

[242] ISO/IEC 14496-4:2004, “Information technology — Coding of audio-visual objects — Part 4: Conformance testing (2nd Ed.)”, 2004.

[243] ISO/IEC 14496-5:2001, “Information technology — Coding of audio-visual objects — Part 5: Reference software (2nd Ed.)”, 2001.

[244] ISO/IEC 14496-6:2000, “Information technology — Coding of audio-visual objects — Part 6: Delivery Multimedia Integration Framework (DMIF) (2nd Ed.)”, 2001.

[245] ISO/IEC TR 14496-7:2004, “Information technology — Coding of audio-visual objects — Part 7: Optimized reference software for coding of audio-visual objects (2nd Ed.)”, 2004.

[246] ISO/IEC 14496-8:2004, “Information technology — Coding of audio-visual objects — Part 8: Carriage of ISO/IEC 14496 contents over IP networks (1st Ed.)”, 2004.

[247] ISO/IEC TR 14496-9:2009, “Information technology — Coding of audio-visual objects — Part 9: Reference hardware description (3rd Ed.)”, 2009.

[248] ISO/IEC 14496-11:2005, “Information technology — Coding of audio-visual objects — Part 11: Scene description and application engine (1st Ed.)”, 2005.

[249] ISO/IEC 14496-12:2008, “Information technology — Coding of audio-visual objects — Part 12: ISO base media file format (3rd Ed.)”, 2008.

[250] ISO/IEC 14496-12:2008/Amd 1:2009, “General improvements including hint tracks, metadata support and sample groups (1st Ed.)”, 2009.

[251] ISO/IEC 14496-13:2004, “Information technology — Coding of audio-visual objects — Part 13: Intellectual Property Management and Protection (IPMP) extensions (1st Ed.)”, 2004.

[252] ISO/IEC 14496-14:2003, “Information technology — Coding of audio-visual objects — Part 14: MP4 file format (1st Ed.)”, 2003.

[253] ISO/IEC 14496-14:2003/FPDAmd 1, “Handling of MPEG-4 audio enhancement layers (1st Ed.)”, 2003.

[254] ISO/IEC 14496-15:2010, “Information technology — Coding of audio-visual objects — Part 15: Advanced Video Coding (AVC) file format (2nd Ed.)”, 2010.

[255] ISO/IEC 14496-15:2010/Amd 1:2011, “Sub-track definitions (1st Ed.)”, 2011.

[256] ISO/IEC 14496-16:2011, “Information technology — Coding of audio-visual objects — Part 16: Animation Framework eXtension (AFX) (3rd Ed.)”, 2011.

[257] ISO/IEC 14496-16:2011/Amd 1:2011, “Efficient representation of 3D meshes with multiple attributes (1st Ed.)”, 2011.

[258] ISO/IEC 14496-17:2006, “Information technology — Coding of audio-visual objects — Part 17: Streaming text format (1st Ed.)”, 2006.

[259] ISO/IEC 14496-18:2004, “Information technology — Coding of audio-visual objects — Part 18: Font compression and streaming (1st Ed.)”, 2004.

[260] ISO/IEC 14496-19:2004, “Information technology, “Coding of audio-visual objects — Part 19: Synthesized texture stream (1st Ed.)”, 2004.

[261] ISO/IEC 14496-20:2008, “Information technology — Coding of audio-visual objects — Part 20: Lightweight Application Scene Representation (LASeR) and Simple Aggregation Format (SAF) (2nd Ed.)”, 2008.

[262] ISO/IEC 14496-20:2008/Amd 3:2010, “Presentation and Modification of Structured Information (PMSI) (1st Ed.)”, 2010.

[263] ISO/IEC 14496-21:2006, “Information technology — Coding of audio-visual objects — Part 21: MPEG-J Graphics Framework eXtensions (GFX) (1st Ed.)”, 2006.

[264] ISO/IEC 14496-22:2009, “Information technology — Coding of audio-visual objects — Part 22: Open Font Format (2nd Ed.)”, 2009.

[265] ISO/IEC 14496-22:2009/Amd 1:2010, “Support for many-to-one range mappings (1st Ed.)”, 2010.

[266] ISO/IEC 14496-23:2008, “Information technology — Coding of audio-visual objects — Part 23: Symbolic Music Representation (1st Ed.)”, 2008.

[267] ISO/IEC TR 14496-24:2008, “Information technology — Coding of audio-visual objects — Part 24: Audio and systems interaction (1st Ed.)”, 2008.

[268] ISO/IEC 14496-25:2011, “Information technology — Coding of audio-visual objects — Part 25: 3D Graphics Compression Mode (2nd Ed.)”, 2011.

[269] ISO/IEC 14496-26:2010, “Information technology — Coding of audio-visual objects — Part 26: Audio conformance (1st Ed.)”, 2010.

[270] ISO/IEC 14496-26:2010/Amd 2:2010, “BSAC conformance for broadcasting (1st Ed.)”, 2010.

[271] ISO/IEC 14496-27:2009, “Information technology — Coding of audio-visual objects — Part 27: 3D Graphics conformance (1st Ed.)”, 2009.

[272] ISO/IEC 14496-27:2009/Amd 3:2011, “Scalable complexity 3D mesh coding conformance in 3DGCM (1st Ed.)”, 2011.

[273] ISO/IEC CD 14496-28:2011, “Information technology — Coding of audio-visual objects — Part 28: Composite font representation (1st Ed.)”, 2011.

[274] Touradj Ebrahimi, and Caspar Horne, “MPEG-4 natural video coding - an overview”, Signal Processing: Image Communication, Volume 15, Issues 4-5, Pages 365-385, January 2000.

[275] Fernando Pereira, and Touradj Ebrahimi (Editors), “The MPEG-4 Book”, Prentice Hall, ISBN 0130616214, July 2002.

[276] Rob Koenen, “Overview of the MPEG-4 Standard”, ISO/IEC JTC1/SC29/WG11, Document N2725, March MPEG meeting, 1999.

[277] Aggelos K. Katsaggelos, Lisimachos P. Kondi, Fabian W. Meir, Jorn Ostermann, and Guido M. Schuster, “MPEG-4 and Rate-Distortion-Based Shape-Coding Techniques”, Proceedings of the IEEE, Volume 86, Issue 6, Pages 1126-1154, June 1998.

[278] Iain E. Richardson, “H.264 and MPEG-4 Video Compression Standard: Video Coding for Next-generation Multimedia”, John Wiley & Sons, ISBN 0470848375, September 2003.

[279] Iain E. Richardson, “The H.264 Advanced Video Compression Standard (2nd Edition)”, John Wiley & Sons, ISBN 0470516925, August 2010.

[280] Yun Q. Shi, and Huifang Sun, “Image and video compression for multimedia engineering: fundamentals, algorithms, and standards (2nd Edition)”, CRC Press, ISBN 9780849373640, 2008.

[281] Lajos L. Hanzo, Peter J. Cherriman, and Jurgen Streit, “Video Compression and Communications: From Basics to H.261, H.263, H.264, MPEG4 for DVB and HSDPA-Style Adaptive Turbo-Transceivers (2nd Edition)”, John Wiley & Sons, ISBN 9780470519929, September 2007.

[282] Jae-Beom Lee, and Hari Kalva, “The VC-1 and H.264 Video Compression Standards for Broadband Video Services”, Springer Publishing Company, ISBN 0387710426, September 2008.

[283] Jie Dong, and King Ngi Ngan, “Present and future video coding standards”, Chang Wen Chen, Zhu Li, and Shiguo Lian (Editors), “Intelligent Multimedia Communication: Techniques and Applications”, Springer-Verlag Publisher, ISBN 9783642116858, Pages 75-124, January 2010.

[284] Garry J. Sullivan, Pankaj N. Topiwala, and Ajay Luthra, “The H.264/AVC advanced video coding standard: overview and introduction to the fidelity range extensions”, Proceedings of SPIE Volume 5558: Conference on Applications of Digital Image Processing XXVII, Pages 454-474, 2004.

[285] Detlev Marpe, Thomas Wiegand, and Gary J. Sullivan, “The H.264/MPEG4 Advanced Video Coding Standard and its Applications”, IEEE Communications Magazine, Volume 44, Issue 8, Pages 134-143, August 2006.

[286] Soon-kak Kwon, A. Tamhankar, and K. R. Rao, “Overview of H.264/MPEG-4 part 10”, Journal of Visual Communication and Image Representation, Volume 17, Issue 2, Pages 186-216, April 2006.

[287] Thomas Wiegand, Garry J. Sullivan, Gisle Bjontegaard, and Ajay Luthra, “Overview of the H.264/AVC video coding standard”, IEEE Transactions on Circuits Systems for Video Technology, Volume 13, Issue 7, Pages 560-576, July 2003.

[288] Gary J. Sullivan, Haoping Yu, Shun-ichi Sekiguchi, Huifang Sun, Thomas Wedi, Steffen Wittmann, Yung Lyul Lee, C. Andrew Segall, and Teruhiko Suzuki, “New standardized extensions of MPEG4-AVC/H.264 for professional quality video applications”, Proceedings of the 2007 IEEE International Conference on Image Processing (ICIP2007), San Antonio, U.S.A., Volume 1, Pages 13-16, September 2007.

[289] Heiko Schwarz, Detlev Marpe, and Thomas Wiegand, “Overview of the Scalable H.264/MPEG4-AVC Extension”, Proceedings of the 2006 IEEE International Conference on Image Processing (ICIP 2006), Atlanta, U.S.A., Pages 161-164, October 2006.

[290] Thomas Wiegand, Gary Sullivan, Julien Reichel, Heiko Schwarz, and Mathias Wien, "Joint Draft 5: Scalable Video Coding", ISO/IEC JTC1/SC29/WG11 and ITU-T SG16 Q.6, Document JVT-R201, 18th Meeting: Bangkok, Thailand, 14-20 January, 2006.

[291] Anthony Vetro, Purvin Pandit, Hideaki Kimata, Aljoscha Smolic, and Ye-Kui Wang, "Joint Draft 8.0 on Multiview Video Coding", ISO/IEC JTC1/SC29/WG11 and ITU-T SG16 Q.6, Document JVT-AB204, 28th Meeting: Hannover, DE, 20-25 July, 2008.

[292] Karsten Müller, Philipp Merkle, Heiko Schwarz, Tobias Hinz, Aljoscha Smolic, and Thomas Wiegand, "Multi-View Video Coding Based on H.264/AVC Using Hierarchical B-Frames", Proceedings of the 25th Picture Coding Symposium (PCS 2006), Beijing, China, Pages 385-390, April 2006.

[293] Anthony Vetro, Purvin Pandit, Hideaki Kimata, and Aljoscha Smolic, "Joint Multiview Video Model (JMVM) 5.0", ISO/IEC JTC1/SC29/WG11 and ITU-T SG16 Q.6, Document JVT-X207, 24th Meeting: Geneva, CH, 29 June – 5 July, 2007.

[294] Jan De Cock, "Compressed-Domain Transcoding of H.264/AVC and SVC Video Streams", Ph.D Dissertation, Universiteit Gent, September 2009.

[295] Gary Sullivan, Tom McMahon, Thomas Wiegand, Detlev Marpe, and Ajay Luthra, "Draft text of H.264/AVC Fidelity Range Extensions Amendment", ISO/IEC JTC1/SC29/WG11 and ITU T SG16 Q.6, Document JVT-L047, 12th Meeting: Redmond, WA, USA 17-23 July, 2004.

[296] Kuo-Liang Chung, and Lung-Chun Chang, "A new predictive search area approach for fast block motion estimation", IEEE Transactions on Image Processing, Volume 12, Issue 6, Pages 648-652, June 2003.

[297] Thomas Wiegand, Gary Sullivan, Julien Reichel, Heiko Schwarz, and Matias Wien, "Joint Draft ITU-T Rec. H.264 | ISO/IEC 14496-10 / Amd.3 Scalable video coding", ISO/IEC JTC1/SC29/WG11 and ITU-T SG16 Q.6, Document JVT-X201, 24th Meeting: Geneva, CH, 29 June – 5 July, 2007.

[298] Thomas Wiegand, Heiko Schwarz, Anthony Joch, Faouzi Kossentini, and Gary J. Sullivan, "Rate-Constrained Coder Control and Comparison of Video Coding Standards", IEEE Transactions on Circuits and Systems for Video Technology, Volume 13, Issue 7, Pages 688-703, July 2003.

[299] Joint Video Team (JVT) of ISO/IEC MPEG and ITU-T VCEG, "Joint Final Committee Draft (JFCD) of Joint Video Specification (ITU-T Rec. H.264 | ISO/IEC 14496-10 AVC)", ISO/IEC JTC1/SC29/WG11 and ITU-T SG16 Q.6., Document JVT-D157, 4th Meeting: Klagenfurt, Austria, 22-26 July, 2002

[300] Alois M. Bock, "Video Compression Systems: From first principles to concatenated codecs", The Institution of Engineering and Technology, ISBN 0863419631, July 2009.

[301] W. K. Cham, "Family of order-4 four-level orthogonal transforms", Electronic Letters, Volume 19, Issue 21, Pages 869–871, October 1983.

[302] Antti Hallapuro, Marta Karczewicz, and Henrique Malvar, "Low complexity transform and quantisation—Part I: Basic implementation", ISO/IEC JTC1/SC29/WG11 and ITU-T SG16 Q.6, Document JVT-B38, 2nd Meeting: Geneva, CH, Jan. 29 - Feb. 1, 2002.

[303] Peter List, Anthony Joch, Jani Lainema, Gisle Bjøntegaard, and Marta Karczewicz, "Adaptive deblocking Filter", IEEE Transactions on Circuits and Systems for Video Technology, Volume 13, Issue 7, Pages 614–619, July 2003.

[304] Simone Milani, "Source and Joint Source-Channel Coding for Video Transmission over Lossy Networks", Ph.D Dissertation, Università degli Studi di Padova 2007.

[305] Detlev Marpe, Heiko Schwarz, and Thomas Wiegand, "Context-based adaptive binary arithmetic coding in the H.264/AVC video compression standard", IEEE Transaction on Circuits Systems and Video Technology, Volume 13, Issue 7, Pages 620–636, July 2003.

[306] Marios C. Angelides, and Harry Agius, "The handbook of MPEG applications: standards in practice", Wiley InterScience, ISBN 0470750073, January 2011.

[307] Haohong Wang, Lisimachos Kondi, Ajay Luthra, and Song Ci, "4G Wireless Video Communications", Wiley InterScience, ISBN 9780470773079, April 2009.

[308] MPEG Video, "MPEG-4 Video Verification Model 8.0", ISO/IEC JTC1/SC29/WG11, Document N1796, Stockholm MPEG meeting, July 1997.

[309] Gary J. Sullivan, and Thomas Wiegand, "Video Compression—From Concepts to the H.264/AVC Standard", Proceedings of the IEEE, Volume 93, Issue 1, Pages 18- 31, January 2005.

[310] Jordi Ribas-Corbera, Philip A. Chou, and Shankar L. Regunathan, "A generalized hypothetical reference decoder for H.264/AVC", IEEE Transactions on Circuits and Systems for Video Technology, Volume 13, Issue 7, Pages 674–687, July 2003.

[311] Thomas Wiegand, "Study of Final Committee Draft of Joint Video Specification (ITU-T Rec. H.264 | ISO/IEC 14496-10 AVC), Draft 2", ISO/IEC JTC1/SC29/WG11 and ITU-T SG16 Q.6, Document JVT-F100d2, 6th Meeting: Awaji, Island, JP, 5-13 December, 2002.

[312] Eric Viscito, "HRD and Related Issues", ISO/IEC JTC1/SC29/WG11 and ITU-T SG16 Q.6, Document JVT-D131, Klagenfurt, Austria, 22-26 July, 2002.

[313] Gary Sullivan, and Heiko Schwarz, "Editing state of text relating to ITU-T Rec. H.264 | ISO/IEC 14496-10 Amendments 1 and 2", ISO/IEC JTC1/SC29/WG11 and ITU-T SG16 Q.6, Document JVT-X066, 24th Meeting: Geneva, CH, 29 June – 5 July, 2007.

[314] Lujun Yuan, Wen Gao, and Yan Lu, "An Improved HRD Model for JVT Standard", ISO/IEC JTC1/SC29/WG11 and ITU-T SG16 Q.6, Document JVT-E091, 5th Meeting: Geneva, Switzerland, 9-17 October, 2002.

[315] Paulo Nunes, "Rate Control for Object-based Video Coding", Ph.D Dissertation, Instituto Superior Técnico, July 2007.

[316] ITU-T Recommendation H.264.1, "Conformance specification for H.264 advanced video coding", Geneva, Switzerland, June 2008.

[317] ITU-T and ISO/IEC JTC 1, "Reference software for advanced video coding", ITU-T Rec. H.264.2 & ISO/IEC 14496-5 (MPEG-4 Reference Software), 2005 (latest approved version). Available from: http://www.itu.int/rec/T-REC-H.264.2 Draft versions are available for download from http://iphome.hhi.de/suehring/tml/. Retrieved 31 January 2012.

[318] Jinhyun Cho, Soonwoo Choi, and Soo-Ik Chae, "Constrained-Random Bitstream Generation for H.264/AVC Decoder Conformance Test", IEEE Transactions on Consumer Electronics, Volume 56, Issue 2, Pages 848-855, May 2010.

[319] Jordi Ribas-Corbera, and Shaw-Min Lei, "Rate control for low-delay video communications", ITU Study Group 16, Video Coding Experts Group, Document Q15-A-20, Portland, USA, 24-27 June 1997.

[320] Jordi Ribas-Corbera, and Shaw-Min Lei, "A Frame-Layer Bit Allocation for H.263+", IEEE Transactions on Circuits and Systems for Video Technology, Volume 10, Issue 7, Pages 1154-1158, October 2000.

[321] Guy Côté, Berna Erol, Michael Gallant, and Faouzi Kossentini, "H.263+: Video Coding al Low Bit Rates", IEEE Transactions on Circuits and Systems for Video Technology, Volume 8, Issue 7, Pages 849-866, November 1998.

[322] ITU-T SG16/Q15, "Video Codec Test Model, Near-Term, Version 10 (TMN10)", Document Q15-D-65, Tampere, Finland, April 1998.

[323] Thomas Wiegand, and Bernd Girod, "Lagrange multiplier selection in hybrid video coder control", Proceedings of the 2001 IEEE International Conference on Image Processing (ICIP 2001), Volume 3, Pages 542–545, October 2001.

[324] Heiko Schwarz, and Thomas Wiegand, "MPEG-4 anchors for the MPEG Call For Proposals On New Tools For Video Compression Technology", ITU Study Group 16, Video Coding Experts Group, Document VCEG-M49, March 2001.

[325] Joint Video Team (JVT) of ISO/IEC MPEG and ITU-T VCEG, "Working Draft Issue 2, Revision 2 (WD-2)", Document JVT-B118R2, Geneva, Switzerland, January 29-February 1, 2002.

[326] Paulo Nunes, and Fernando Pereira, "Object-based rate control for the MPEG-4 visual simple profile", Proceedings of the Second International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS 1999), Berlin, Germany, Volume 1, Pages 161-164, May 1999.

[327] MPEG Video, "MPEG-4 Video Verification Model 18.0", Doc. ISO/IEC JTC1/SC29/WG11 N1796, Pisa MPEG meeting, January 2001.

[328] Hung-Ju Lee, Tihao Chiang, and Ya-Qin Zhang, "Scalable rate control for MPEG-4 video", IEEE Transaction on Circuits and Systems for Video Technology, Volume 10, Issue 6, Pages 878-894, September 2000.

[329] Tihao Chiang, Ya-Qin Zhang, J. Lee, S. Martucci, H. Peterson, I. Sodagar, R. Suryadevara, and C. Wine, "A Rate Control Scheme Using A New Rate-Distortion Model", ISO/IEC JTC1/SC29/WG11, Document M0436, Dallas MPEG meeting, November 1995.

[330] Hung-Ju Lee, Tihao Chiang, and Ya-Qin Zhang, "Scalable rate control for very low bitrate video", Proceedings of the 1997 IEEE International Conference on Image Processing (ICIP 1997), Volume 2, Pages 768–771, October 1997.

[331] Zongze Wu, Shengli Xie, Kexin Zhang and Rong, Wu, "Rate Control in Video Coding", Javier Del Ser Lorente (Editor), "Recent Advances on Video Coding", InTech, ISBN 9789533071817, Pages 79-116, 2011.

[332] Tihao Chiang, Iraj Sodagar, Stephen Martucci, and Ya-Qin Zhang, "Status report on core experiment Q2 improved rate control", ISO/IEC JTC1/SC29/WG11, Document M1109, Tempere MPEG meeting, July 1996.

[333] Sung-Gul Ryoo, Seong-Jin Kim, and Yang-Seock Seo, "Rate Control Tool: Based on Human Visual Sensitivity (HVS) for Low Bitrate Coding", ISO/IEC JTC1/SC29/WG11, Document MPEG96/0566, Munchen MPEG meeting, Germany, January 1996.

[334] Anthony Vetro, and Huifang Sun, "CE Q2: Multiple video object rate control", ISO/IEC JTC1/SC29/WG11, Document ISO/IEC M2219, Stockholm MPEG meeting, July 1997.

[335] Hung-Ju Lee, Tihao Chiang, and Ya-Qin Zhang, "Multiple-VO rate control and B-VO rate control", ISO/IEC JTC1/SC29/WG11, Document M2554, Stockholm MPEG meeting, July 1997.

[336] Paulo Nunes, and Luis Ducla Soares, "Rate Control and Error Resilience for Object-Based Video Coding", Chang Wen Chen, Zhu Li, and Shiguo Lian (Editors), "Intelligent Multimedia Communication: Techniques and Applications", Springer-Verlag Publisher, ISBN 9783642116858, Pages 1-50, January 2010.

[337] Xiaodong Cai, Falah H. Ali, and Elias Stipidis, “Object-based video coding with dynamic quality control”, Image and Vision Computing, Volume 28, Issue 3, Pages 285–297, March 2010.

[338] Paulo Nunes, and Fernando Pereira, “Joint Rate Control Algorithm for Low-Delay MPEG-4 Object-Based Video Encoding”, IEEE Transactions on Circuits and Systems for Video Technology, Volume 19, Issue 9, Pages 1274-1288, September 2009.

[339] Zhengguo Li, Feng Pan, Keng Pang Lim, Genan Feng, Xiao Lin, and Susanto Rahardja, “Adaptive basic unit layer rate control for JVT”, ISO/IEC JTC1/SC29/WG11 and ITU-T SG16 Q.6, Document JVT-G012r1, 7th Meeting: Pattaya II ,Thailand, 7-14 March, 2003.

[340] Siwei Ma, Wen Gao, Feng Wu, and Yan Lu, “Rate control for JVT video coding scheme with HRD considerations”, Proceedings of the 2003 IEEE International Conference on Image Processing (ICIP 2003), Barcelona, Spain, Volume 3, Pages 793–796, September 2003.

[341] Yuan Wu, Lin Shouxun, Zhang Yongdong, Luo Haiyong, and Yuan Wen, “Optimum Bit Allocation and Rate Control for H.264/AVC”, ISO/IEC JTC1/SC29/WG11 and ITU-T SG16 Q.6, Document JVT-O016, 15th Meeting, Busan, South Korean, 16-22 April, 2005.

[342] Z. G. Li, F. Pan, K. P. Lim, X. Lin, and S. Rahardja, “Adaptive rate control for H.264”, Proceedings of the 2004 IEEE International Conference on Image Processing (ICIP 2004), Volume 2, Pages 745-748, October 2004.

[343] Z. G. Li, F. Pan, K. P. Lim, G. N. Feng, S. Rahardja and D. J. Wu, “Adaptive frame layer rate control for H.264”, Proceedings of the 2003 IEEE International Conference on Multimedia and Expo (ICME 2003), Volume 1, Pages 581-584, June 2003.

[344] Siwei Ma, Wen Gao, and Yan Lu, “Rate Control on JVT Standard”, ISO/IEC JTC1/SC29/WG11 and ITU-T SG16 Q.6, Document JVT-D030, 4th Meeting, Klagenfurt, Austria, 22-26 July, 2002.

[345] Kannan Ramchandran, Antonio Ortega and Martin Vetterli, “Bit allocation for dependent quantisation with applications to MPEG video codec”, Proceedings of the 1993 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 1993), Minneapolis, Volume 5, Pages 381-384, March 1993.

[346] Gerjan Keesman, Imran Shah, and Rene Klein-Gunnewiek, “Bit-rate control for MPEG encoders”, Signal Processing: Image Communication, Volume 6, Issue 6, Pages 545-560, February 1995.

[347] Aggelos K. Katsaggelos, and Gerry Melnikov, “Rate-Distortion Techniques in Image and Video Coding”, Ling Guan, Sun-Yuan Kung, and Jan Larsen (Editors), “Multimedia Image and Video Processing”, Boca Raton: CRC Press LLC, ISBN 0849334926, 2001.

[348] Toby Berger, “Rate distortion theory: a mathematical basis for data compression”, Prentice-Hall, Inc., Englewood Cliffs, NJ, ISBN 0137531036, 1971.

[349] Claude E. Shannon, “A mathematical theory of communication”, Bell System Technical Journal, Volume 27 (July and October 1948), Pages 379- 423 and 623-656, 1948.

[350] Antonio Ortega, and Kannan Ramchandran, “Rate-Distortion Methods for Image and Video Compression”, IEEE Signal Processing Magazine, Volume 15, Issue 6, Pages 23-50, November 1998.

[351] Allen Gersho, “Asymptotically optimal block quantisation”, IEEE Transactions on Information Theory, Volume 25, Issue 4, Pages 373-380, July 1979.

[352] David L. Neuhoff, ‘The other asymptotic theory of source coding”, Robert Calderbank, G. David Forney, Jr., Nader Moayeri (Editors), “DIMACS Series in Discrete Mathematics and Theoretical Computer Science: Coding and Quantization", American Mathematical Society, Volume 14, Page 55-66, February 1993.

[353] J. J. Y. Huang, and P. M. Schultheiss, “Block quantisation of correlated Gaussian random variables”, IEEE Transactions on Communications Systems, Volume 11, Issue 3, Pages 289–296, 1963.

[354] Yair Shoham, Allen Gersho, “Efficient bit allocation for an arbitrary set of quantisers”, IEEE Transactions on Acoustics, Speech, and Signal Processing, Volume 36, Issue 9, Pages 1445–1453, 1988.

[355] Zhu Li, Aggelos K. Katsaggelos, and Guido M. Schuster, “Rate-Distortion Optimal Video Summarization and Coding”, Yap-Peng Tan, Kim Hui Yap, and Lipo Wang (Editors), “Intelligent Multimedia Processing with Soft Computing”, Springer Berlin Heidelber, ISBN 9783540230533, Pages 171-204, 2005.

[356] Guido M. Schuster, and Aggelos K. Katsaggelos, “Rate-Distortion Based Video Compression: Optimal Video Frame Compression and Object Boundary Encoding”, Kluwer Academic Publishers, ISBN 0792398505, 1996.

[357] G. David Forney Jr., “The Viterbi algorithm”, Proceedings of the IEEE, Volume 61, Issue 3, Pages 268–278, 1973.

[358] Minqiang Jiang, “Adaptive Rate Control for Advanced Video Coding”, Ph.D Dissertation, Santa Clara University, January 2006.

[359] Cheng-Yu Pai, “Rate Control and Constant Quality Rate Control for MPEG Video Compression and Transcoding”, Ph.D Dissertation, Montreal, Quebec, Canada, 2006.

[360] Min Dai, “Rate-Distortion Analysis for Scalable Coders”, Ph.D Thesis, Texas University, December 2004.

[361] Zhenzhong Chen, and King Ngi Ngan, “Linear rate-distortion models for MPEG-4 shape coding”, IEEE Transactions on Circuits and Systems for Video Technology, Volume 14, Issue 6, Pages 869–873, June 2004.

[362] Zhenzhong Chen, and King Ngi Ngan, “Rate-distortion analysis for MPEG-4 binary shape coding”, Proceedings of the 2005 IEEE International Symposium on Intelligent Signal Processing and Communications Systems (ISPACS 2005), Hong Kong, Pages 801-804, 2005.

[363] Hyun Mun Kim, “Adaptive rate control using nonlinear regression”, IEEE Transactions on Circuits and Systems for Video Technology, Volume 13, Issue 5, Pages 432–439, 2003.

[364] Arun N. Netravali and John O. Limb, "Picture coding: A review”, Proceedings of the IEEE, Volume 68, Issue 3, Pages 7-12, March 1960.

[365] Zhihai He, Yong Kwan Kim, and Sanjit K Mitra, “Low-delay rate control for DCT video coding via ρ-domain source modeling”, IEEE Transactions on Circuits and Systems for Video Technology, Volume 11, Issue 8, Pages 928-940, August 2001.

[366] Yong Kwan Kim, Zhihai He, and Sanjit K. Mitra, “A novel linear source model and a unified rate control algorithm for H.263/MPEG-2/MPEG-4”, Proceedings of the 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2001), Salt Lake City, UT, Volume 3, Pages 1777-1780, May 2001.

[367] Nejat Kamaci, Yucel Altunbasak, and Russel M. Mersereau, “Frame bit allocation for the H.264/AVC video coder via Cauchy-density-based rate and distortion models”, IEEE Transactions on Circuits and Systems for Video Technology, Volume 15, Issue 5, Pages 994–1006, August 2005.

[368] Zhenzhong Chen, and King Ngi Ngan, “Recent advances in rate control for video coding”, Signal Processing: Image Communication, Volume 22, Issue 1, Pages 19-38, January 2007.

[369] Zhenzhong Chen, and King Ngi Ngan, “Joint texture-shape optimization for MPEG-4 multiple video objects”, IEEE Transactions on Circuits and Systems for Video Technology, Volume 15, Issue 9, Pages 1170–1174, September 2005.

[370] Zhongwei Zhang, Guizhong Liu, Hongliang Li, and Yongli Li, "A novel PDE-based rate distortion model for rate control", IEEE Transactions on Circuits and Systems for Video Technology, Volume 15, Issue 11, Pages 1354–1364, 2005.

[371] Yui-Lam Chan, and Wan-Chi Siu, "An efficient search strategy for block motion estimation using image features", IEEE Transactions on Image Processing, Volume 10, Issue 8, Pages 1223-1238, August 2001.

[372] A. Puri, H.-M. Hang, and D. L. Schilling, "Interframe coding with variable block-size motion compensation", Proceedings of the IEEE 1987 Global Telecommunications Conference (GLOBECOM 1987), Tokyo, Japan, Pages 65-69, 1987.

[373] Injong Rhee, Graham R. Martin, S. Muthukrishnan, and Roger A. Packwood, "Quadtree-structured variable-size block-matching motion estimation with minimal error", IEEE Transactions on Circuits and Systems for Video Technology, Volume 10, Issue 1, Pages 42-50, February 2001.

[374] Jiann-Jone Chen, and D.W. Lin, "Optimal bit allocation for video coding under multiple constraints", Proceedings of the 1996 IEEE International Conference on Image Processing (ICIP 1996), Lausanne, Switzerland, Volume 3, Pages 403 - 406, 1996.

[375] Zihai He, and Sanjit K. Mitra, "Optimum bit allocation and accurate rate control for video coding via rho-domain source modelling", IEEE Transactions on Circuits and Systems for Video Technology, Volume 12, Issue 10, Pages 840–849, 2002.

[376] Guido M. Schuster, and Aggelos K. Katsaggelos, "A video compression scheme with optimal bit allocation among segmentation motion and residual error", IEEE Transactions on Image Processing, Volume 6, Issue 11, Pages 1487–1502, 1997.

[377] Guido M. Schuster, and Aggelos K. Katsaggelos, "A theory for the optimal bit allocation between displacement vector field and displaced frame difference", IEEE Journal on Selected Areas in Communications, Volume 15, Issue 9, Pages 1739–1751, 1997.

[378] Jianning Zhang, Yuwen He, Shiqiang Yang, and Yuzhuo Zhong, "Performance and complexity joint optimization for H.264 video coding", Proceedings of the 2003 IEEE International Symposium on Circuits and Systems (ISCAS 2003), Volume 2, Pages 888-891, May 2003.

[379] Hye-Yeon Cheng Tourapis, and Alexis M. Tourapis, "Fast motion estimation within the H.264 codec", Proceedings of the 2003 IEEE International Conference on Multimedia and Expo (ICME 2003), Volume 3, Pages 517-520, July 2003.

[380] Jordis Ribas-Corbera, and David L. Neuhoff, "Optimizing block size in motion compensated video coding", Journal of Electronic Imaging, Volume 7, Number 1, Pages 155-165, January 1998.

[381] Zhenzhong Chen, and K. N. Ngan, "Optimal bit allocation for MPEG-4 multiple video objects", Proceedings of the 2004 IEEE International Conference on Image Processing (ICIP 2004), Singapore, Volume 2, Pages 761-764, October 2004.

[382] Minqiang Jiang, Xiaoquan Yi, and Nam Ling, "Improved Frame-Layer Rate Control for H.264 Using MAD Ratio", Proceedings of the 2004 IEEE International Symposium on Circuits and Systems (ISCAS 2004), Vancouver, Canada, Volume 3, Pages 813–816, May 2004.

[383] Minqiang Jiang, Xiaoquan Yi, and Nam Ling, "Frame Layer Bit Allocation Scheme for Constant Quality Video", Proceedings of the 2004 IEEE International Conference on Multimedia and Expo (ICME 2004), Taipei, Taiwan, TP4-4, Volume 2, Pages 1055-1058, June 2004.

[384] Xiaoquan Yi, and Nam Ling, "Rate Control Using Enhanced Frame Complexity Measure for H.264 Video", Proceedings of the 2004 IEEE Workshop on Signal Processing Systems (SiPS), Austin, U.S.A., Pages 263 – 268, October 2004.

[385] Xiaoquan Yi, and Nam Ling, "Improved H.264 Rate Control by Enhanced MAD-Based Frame Complexity Prediction", Journal of Visual Communication and Image Representation (Special Issue on Emerging H.264/AVC Video Coding Standard), Volume 17, Issue 2, Pages 407- 424, April 2006.

[386] Zhao Min, Xin Jin, and Satoshi Goto, "Novel Real-time Rate Control Algorithm for Constant Quality H.264/AVC High Vision Codec", Proceedings of the 5th International Colloquium on Signal Processing & Its Applications (CSPA 2009), Pages 323-326, March 2009.

[387] Do-Kyoung Kwon, Mei-Yin Shen, and C. C. Jay Kuo, "Rate control for H.264 video with enhanced rate and distortion models", IEEE Transactions on Circuits and Systems for Video Technology, Volume 17, Issue 5, Pages 517–529, May 2007.

[388] Jianpeng Dong, and Nam Ling, "On Model Parameter Estimation for H.264/AVC Rate Control", Proceedings of the 2007 IEEE International Symposium on Circuits and Systems (ISCAS 2007), Volume 1, Pages 289-292, May 2007.

[389] Mohammed Golam Sarwer, Lai Man Po, and Q. M. Jonathan Wu, "Bit Rate Estimation for Cost Function of H.264/AVC", Kazuki Nishi (Editor), "Multimedia", InTech, ISBN 9789537619879, Pages 257-280, February 2010.

[390] Bojun Meng, Oscar Au, Chi-wah Wong, and Hong-Kwai Lam, "Efficient intra-prediction mode selection for 4x4 blocks in H.264", Proceeding of 2003 IEEE International Conference on Multimedia and Expo (ICME 2003), Baltimore, U.S.A., Volume 3, Pages 521-524, 2003.

[391] Chun-Ling Yang, Lai Man Po, and Wing-Hong Lam, "A fast H.264 Intra Prediction algorithm using macroblock properties", Proceedings of the 2004 IEEE International Conference on Image Processing (ICIP 2004), Volume 1, Pages 461-464, 2004.

[392] Feng Pan, Xiao Lin, Susanto Rahardja, Keng Pang Lim, Z. G. Li, Dajun Wu, and Si Wu, " Fast Mode Decision Algorithm for Intra-prediction in H.264/AVC Video Coding", IEEE Transactions on Circuits and Systems for Video Technology, Volume 15, Issue 7, Pages 813- 822, July 2005.

[393] Chao-Hsuing Tseng, Hung-Ming Wang, and Jar-Ferr Yang, "Enhanced Intra 4x4 Mode Decision for H.264/AVC Coder", IEEE Transaction on Circuits and Systems for Video Technology, Volume 16, Issue 8, Pages 1027-1032, August 2006.

[394] Shuijiong Wu, Peilin Liu, Yiqing Huang, Qin Liu, and Takeshi Ikenaga, "Constant Bit-Rate Multi-Stage Rate Control for Rate-Distortion Optimized H.264/AVC Encoders", IEICE Transactions on Information and Systems, Volume E93-D, Issue 7, Pages 1716-1726, July 2010.

[395] Shuijiong Wu, Yiqing Huang, Qin Liu, and Takeshi Ikenaga, "Rate-Distortion Optimized Multi-Stage Rate Control Algorithm for H.264/AVC Video Coding", Proceedings of the 17th European Signal Processing Conference (EUSIPCO 2009), Glasgow, Scotland, Pages 1809-1813, August 2009.

[396] Shuijiong Wu, Peilin Liu, Yiqing Huang, Qin Liu, and Takeshi Ikenaga, "On bit allocation and Lagrange Multiplier adjustment for rate-distortion optimized H.264 rate control", Proceedings of the 2009 IEEE International Workshop on Multimedia Signal Processing (MMSP 2009), Pages 1-6, October 2009.

[397] Lulin Chen, and Ilie Garbacea, "Adaptive Lambda estimation in Lagrangian rate-distortion optimization for video coding", Proceedings of SPIE Volume 6077: Visual Communications and Image Processing 2006, Pages 60772B 1–8, January 2006.

[398] Xiang Li, Norbert Oertel, Andreas Hutter, and André Kaup, "Laplace Distribution Based Lagrangian Rate Distortion Optimization for Hybrid Video Coding", IEEE Transactions on Circuits and Systems for Video Technology, Volume19, Issue 2, Pages 193-205, February 2009.

[399] Xiang Li, Peter Amon, Andreas Hutter, and André Kaup, "One-pass multi-layer rate-distortion optimization for quality scalable video coding", Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2009), Pages 637-640, April 2009.

[400] Miohui Wang, and Bo Yan, "Lagrangian multiplier based joint three-layer rate control for H.264/AVC", IEEE Signal Processing Letters, Volume 16, Issue 8, Pages 679–682, August 2009.

[401] Minqiang Jiang, and Nam Ling, "On Lagrange multiplier and quantizer adjustment for H.264 frame-layer video rate control", IEEE Transactions on Circuits and Systems for Video Technology, Volume 16, Issue 5, Pages 663–669, May 2006.

[402] Jun Zhang, Xiaoquan Yi, Nam Ling, and Weija Shang, "Context adaptive Lagrange multiplier (CALM) for rate-distortion optimal motion estimation in video coding", IEEE Transactions on Circuits and Systems for Video Technology, Volume 20, Issue 6, Pages 820–828, June. 2010.

[403] En-hui Yang, and Xiang Yu, "Soft Decision Quantization for H.264 With Main Profile Compatibility", IEEE Transactions on Circuits and Systems for Video Technology, Volume 19, Issue 1, Pages 122-127, January 2009.

[404] En-hui Yang, and Xiang Yu, "Rate Distortion Optimization of H.264 with Main Profile Compatibility", Proceedings of the 2006 IEEE International Symposium on Information Theory, Pages282-286, July 2006.

[405] En-Hui Yang, and Xiang Yu, "Rate Distortion Optimization for H.264 Interframe Coding: A General Framework and Algorithms", IEEE Transactions on Image Processing, Volume 16, Issue 7, Pages 1774-1784, July 2007.

[406] Marta Karczewicz, Peisong Chen, Yan Ye, and Rajan Joshi, "R-D based quantization in H.264", Proceedings of SPIE Volume 7443: Applications of Digital Image Processing XXXII, Pages 744314-744314-8, 2009.

[407] Marta Karczewicz, Yan Ye, and Insuk Chong, "Rate distortion optimized quantization", ITU-T Q.6/SG16 VCEG, VCEG-AH21, 34th Meeting, Antalya Turkey, January 2008.

[408] Fu-Chuang Chen, and Yi-Pin Hsu, "Rate-Distortion Optimization of H.264/AVC Rate Control with Novel Distortion Prediction Equation", IEEE Transactions on Consumer Electronics, Volume 57, Issue 3, Pages 1264-1270, August 2011.

[409] Xin Zhao, Li Zhang, Siwei Ma, and Wen Gao, "Video Coding with Rate-Distortion Optimized Transform", IEEE Transactions on Circuits and Systems for Video Technology, Volume 22, Issue 1, January 2012.

[410] Reference ITU-T VCEG-KTA Software, ITU-T VCEG, version 2.6r1. Available from: http://iphome.hhi.de/suehring/tml/download/KTA. Retrieved 31 January 2012.

[411] X. K. Yang, W. S. Lin, Z. K. Lu, E. P. Ong, and S. S. Yao, "Just noticeable distortion model and its applications in video coding", Signal Processing: Image Communication, Volume 20, Issue 7, Pages 662–680, August 2005.

[412] Xiaokang Yang, Weisi Lin, Zhongkhang Lu, EePing Ong, and Susu Yao, "Motion-compensated residue pre-processing in video coding based on just-noticeable distortion profile", IEEE Transactions on Circuits and Systems for Video Technology, Volume 15, Issue 6, Pages 742–752, June 2005.

[413] Zhongjie Zhu, Yuer Wang, Yongqiang Bai, and Gangyi Jiang, "On Optimizing H. 264/AVC Rate Control by Improving R-D Model and Incorporating HVS Characteristics", EURASIP Journal on Advances in Signal Processing, Volume 2010, Article ID 830605, Pages 1-10, February 2010.

[414] Zhanzhong Chen, and Christine Guillemot, "Perceptually-friendly H.264/AVC video coding based on foveated just-noticeable-distortion model", IEEE Transactions on Circuits and Systems for Video Technology, Volume 20, Issue 6, Pages 806–819, June 2010.

[415] Zhicheng Li, Shiyin Qin, and Laurent Itti, "Visual attention guided bit allocation in video compression", Image and Vision Computing, Volume 29, Issue 1, Pages 1- 14, 2011.

[416] Matteo Naccari, and Fernando Pereira, "Advanced H.264/AVC-Based Perceptual Video Coding: Architecture, Tools, and Assessment", IEEE Transactions on Circuits and Systems for Video Technology, Volume 21, Issue 6, Pages 766-782, June 2011.

[417] Shiqi Wang, Abdul Rehman, Zhou Wang, Siwei Ma, and Wen Gao, "SSIM-motivated rate distortion optimization for video coding", IEEE Transactions on Circuits and Systems for Video Technology, Volume 22, Issue 4, Pages 516-529, April 2012.

[418] Zhi-Yi Mai, Chun-Ling Yang, and Sheng-Li Xie, "A Novel Rate-Distortion Optimization Based on Structural Similarity in Color Image Encoder", International Journal of Information Technology, Volume 11, Issue 7, Pages 71-80, 2005.

[419] Babu Aswathappa, and K. R. Rao, "Rate-Distortion Optimization using Structural Information in H.264 strictly Intra-frame Encoder", Proceedings of the 42nd South Eastern Symposium on System Theory, Texas, U.S.A., Pages 367-370, March 2010.

[420] Chun-Ling Yang, Rong-Kun Leung, Lai-Man Po, and Zhi-Yi Mai, "An SSIM-optimal H.264/AVC inter frame encoder", Proceedings of the 2009 IEEE International Conference on Intelligent Computing and Intelligent Systems (ICIS 2009), Volume 4, Pages 291-295, November 2009.

[421] Luís Teixeira, "Rate-distortion Analysis for H.264/AVC Video Modelling", Javier Del Ser Lorente (Editor), "Recent Advances on Video Coding", InTech, ISBN 9789533071817, Pages 117-140, 2011.

[422] Yi-Hsin Huang, Tao-Sheng Ou, Po-Yen Su, and Homer H. Chen, "Perceptual Rate-Distortion Optimization Using Structural Similarity Index as Quality Metric", IEEE Transactions on Circuits and Systems for Video Technology, Volume 20, Issue 11, Pages 1614-1624, November 2010.

[423] Homer H. Chen, Yi-Hsin Huang, Po-Yen Su, and Tao-Sheng Ou, "Improving video coding quality by perceptual rate-distortion optimization", Proceedings of the 2010 IEEE International Conference on Multimedia and Expo (ICME 2010), Pages 1287-1292, July 2010.

[424] Tao-Sheng Ou, Yi-Hsin Huang, and Homer H. Chen, "SSIM-Based Perceptual Rate Control for Video Coding", IEEE Transactions on Circuits and Systems for Video Technology, Volume 21, Issue 5, Pages 682 - 691, May 2011.

[425] Shiqi Wang, Abdul Rehman, Zhou Wang, Siwei Ma, and Wen Gao, "Rate-SSIM optimization for video coding", Proceedings of the 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2011), Pages 833-836, 22-27 May 2011.

[426] Abdul Rehman, and Zhou Wang, “Reduced-reference SSIM estimation”, Proceedings of the 2010 International Conference on Image Processing (ICIP 2010), Pages 289–292, September 2010.

[427] Sumohana S. Channappayya, Alan Conrad Bovik, and Robert W. Heath Jr., “Rate bounds on SSIM index of quantized images”, IEEE Transactions on Image Processing, Volume 17, Issue 9, Pages 1624–1639, September 2008.

[428] Zhou Wang, and Alan C. Bovik, “Reduced- and No-Reference Image Quality Assessment”, IEEE Signal Processing Magazine, Volume 28, Issue 6, Pages 29-40, November 2011.

[429] Abdul Rehman, and Zhou Wang, “Reduced-Reference Image Quality Assessment by Structural Similarity Estimation”, IEEE Transactions on Image Processing, Volume 21, Issue 8, Pages 3378-3389, August 2012.

[430] A. Albonico, G. Valenzise, M Naccari, M. Tagliasacchi, and S. Tubaro, “A Reduced-Reference Video Structural Similarity Metric based on No-Reference Estimation of Channel-Induced Distortion”, Proceedings of the 2009 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2009), Pages 1857- 1860, 2009.

[431] Matteo Naccari, Marco Tagliasacchi, and Stefano Tubaro, “No-Reference Video Quality Monotoring for H.264/AVC Coded Video”, IEEE Transactions on Multimedia, Volume 11, Issue 5, Pages 2324-2327, August 2009.

[432] Matteo Naccari, Marco Tagliasacchi, Fernando Pereira, and Stefano Tubaro, “No-Reference Modeling of the Channel induced Distortion at the Decoder for H.264/AVC Video Coding”, Proceedings of the 2008 IEEE International Conference on Image Processing (ICIP 2008), Pages 2324- 2327, 2008.

[433] Patrick Seeling, Martin Reisslein, and Beshan Kulapala, “Network Performance Evaluation with Frame Size and Quality Traces of Single-Layer and Two-Layer Video: A Tutorial”, IEEE Communications Surveys and Tutorials”, Volume 6, Issue 3, Pages 58-78, Third Quarter 2004.

[434] Patrick Seeling, Frank H. P. Fitzek, and Martin Reisslein (Editors), “Video Traces for Network Performance Evaluation - A Comprehensive Overview and Guide on Video Traces and Their Utilization in Networking Research”, Springer Verlag, ISBN 9781402055669, 2007.

[435] Donald Adjeroh, and M. C. Lee, “Scene-adaptive transform domain video partitioning”, IEEE Transaction on Multimedia, Volume 6, Issue 1, Pages 58-69, February 2004.

[436] “Web page of Video Traces Research Group”, Arizona State University, http://trace.eas.asu.edu/yuv/index.html, Retrieved 31 January 2012.

[437] Wes Simpson (Editor), “Video Over IP - IPTV, Internet Video, H.264, P2P, Web TV, and Streaming: A Complete Guide to Understanding the Technology (2nd Edition)”, Focal Press, ISBN 9780240810843, 2008.

[438] Atul Puri, Xuemin Chen, and Ajay Luthra, “Video coding using the H.264/MPEG-4 AVC compression standard”, Signal Processing: Image Communication, Volume 19, Issue 9, Pages 793–849, October 2004.

[439] Hamid Rahim Sheikh, Muhammad Farooq Sabir, and Alan Conrad Bovik , “A Statistical Evaluation of Recent Full Reference Image Quality Assessment Algorithms”, IEEE Transactions on Image Processing, Volume 15, Issue 11, Pages 3440-3451, November 2006.

[440] Stephen B. Vardeman, “Statistics for Engineering Problem Solving”, PWS Publishing Company, ISBN 0534928714, 1994.

[441] A. Y. K. Yan, and M. L. Liou, “Adaptive predictive rate control algorithm for MPEG videos by rate quantisation method”, Proceedings of the 1997 Picture Coding Symposium (PCS 1997), Berlin, Germany, Pages 619-624, September 1997.

[442] Liang-Jin Lin, Antonio Ortega, and C.-C. Jay Kuo, “Rate control using spline-interpolated R-D characteristics”, Proceedings of the 1996 Visual Communications and Image Processing (VCIP 1996), Orlando, U.S.A., Pages 111-122, March 1996.

[443] Emmanuel D. Frimout, Jan Biemond, and Reginald L. Lagendijk, “Forward rate control for MPEG recording”, Proceedings of SPIE Volume 2094: Visual Communication and Image Processing 1993, Pages 184-194, November 1993.

[444] Bo Tao, Heide A. Peterson, and Bradley W. Dickinson, “A rate-quantisation model for MPEG encoders”, Proceedings of the 1997 International Conference in Image Processing (ICIP 1997), Santa Barbara, U.S.A., Volume 1, Pages 338-341, October 1997.

[445] Kyeong Ho Yang, Arnaud Jacquin, and Nikil S. Jayant, “A normalised rate-distortion model for H.263-compatible codecs and its application to quantiser selection”, Proceedings of the 1997 International Conference in Image Processing (ICIP 1997), Santa Barbara, U.S.A., Volume 2, Pages 41-44, October 1997.

[446] Zhou Heng, and Zhao Haiwu, “The Rate Distortion Function of H.264 Transform Coefficients”, ISO/IEC JTC1/SC29/WG11 and ITU-T SG16 Q.6, Document JVT-Q080, 17th Meeting: Nice, FR, 14-21 October, 2005.

[447] Patrick Seeling, and Martin Reisslein, “Video Transport Evaluation With H.264 Video Traces”, IEEE Communications Surveys & Tutorials, Volume 14, Issue 4, Pages 1142-1165, Fourth Quarter 2012.

[448] Sudhir Kumar Srinivasan, Jonathan Vahabzadeh-Hagh, and Martin Reisslein, “The Effects of Priority Levels and Buffering on the Statistical Multiplexing of Single-Layer H.264/AVC and SVC Encoded Video Streams”, IEEE Transaction on Broadcasting, Volume 56, Issue 3, Pages 281-287, September 2010.

[449] Natalia M. Markovich, Astrid Undheim, and Peder J. Emstad, “Classification of slice-based VBR video traffic and estimation of link loss by exceedance”, Computer Networks, Volume 53, Issue 7, Pages 1137–1153, May 2009.

[450] Vladimir Vukadinovic, and Jorg Huschke, “Statistical multiplexing gains of H.264/AVC video in E-MBMS”, Proceedings of the 3rd International Symposium on Wireless Pervasive Computing (ISWPC 2008), Santorini, Greece, Pages 468–474, May 2008.

[451] Patrick Seeling, and Martin Reisslein, “The Rate Variability-Distortion (VD) Curve of Encoded Video and Its Impact on Statistical Multiplexing”, IEEE Transactions Broadcasting, Volume 51, Issue 4, Pages 473–92, December 2005.

[452] Geer Van der Auwera, Prasanth T. David, and Martin Reisslein, “Traffic and quality characterization of single-layer video streams encoded with the H.264/MPEG-4 advanced video coding standard and scalable video coding extension”, IEEE Transactions on Broadcasting, Volume 54, Issue 3, part 2, Pages 698–718, September 2008.

[453] Geert Van der Auwera, Prasanth T. David, and Martin Reisslein, “Traffic characteristics of H.264/AVC variable bit rate video”, IEEE Communications Magazine, Volume 46, Issue 11, Pages 698–718, November 2008.

[454] Astrid Undheim, Yuan Lin, and Peter J. Emstad, “Characterization of slice-based H.264/AVC encoded video traffic”, Proceedings of the 4th European Conference on Universal Multiservice Networks (ECUMN 2007), Toulouse, France, Pages 263–272, February 2007.

[455] Mehdi Rezaei, “Video Streaming over DVB-H”, Fa-Long Luo (Editor), “Mobile Multimedia Broadcasting Standards: Technology and Pratice”, Springer, ISBN 9780387782621, Pages 109-132, 2009.

[456] Martin Fleury, Emmanuel Jammeh, Rouzbeh Razavi, Sandro Moiron, and Mohammed Ghanbari, "Video Streaming in Evolving Networks under Fuzzy Logic Control", Christos J Bouras (Editor), "Trends in Telecommunications Technologies", InTech, ISBN 9789533070728, March 2010.

[457] T. Raghuveera and K. S. Easwarakumar, "An Efficient Statistical Multiplexing Method for H.264 VBR Video Sources for Improved Traffic Smoothing", International Journal of Computer Science and Information Technology, Volume 2, Issue 2, April 2010.

[458] Ahmet Kondoz, "Visual Media Coding and Transmission", John Wiley & Sons, ISBN 9780470740576, 2009.

[459] Sudhir K. Srinivasan, Jonathan Vahabzadeh, and Martin Reisslein, "The Effects of Priority Levels and Buffering on the Statistical Multiplexing of Single-Layer H.264/AVC and SVC Encoded Video Streams", IEEE Transactions on Broadcasting, Volume 56, Issue 3, Pages 281-287, September 2010. (Extended Version) Technical Report, Electrical, Computer, and Energy Engineer, ASU, April 2010.

[460] Geert Van der Auwera, and Martin Reisslein, "Implications of Smoothing on Statistical Multiplexing of H.264/AVC and SVC Video Streams", IEEE Transactions on Broadcasting, Volume 55, Issue 3, Pages 541-558, September 2009.

[461] Peter H. Westerink, R. Rajagopalan, and C. A. Gonzales, "Two-pass MPEG-2 variable-bit-rate encoding", IBM Journal of Research Development 43, Issue 4, Pages 471-488, July 1999.

[462] Luís Teixeira, and Hugo Ribeiro, "Analysis of a two step MPEG video system", Proceedings of the 1997 International Conference on Image Processing (ICIP 1997), Volume 1, Pages 350-352, 1997.

[463] Hai Bing Yin, Xiang Zhong Fang, Li Chen, and Jun Hou, "A practical consistent-quality two-pass VBR video coding algorithm for digital storage application", IEEE Transactions on Consumer Electronics, Volume 50, Issue 4, Pages 1142- 1150, November 2004.

[464] Luís Teixeira, and Artur P. Alves, "MPEG Bitrate Control for Two-Step Coding", Proceedings of the 3rd International Conference Communicating by Image and Multimedia, Bordeaux, France, Pages 327-328, May 1996.

[465] Hai Bing Yin, Xiang Zhong Fang, and Yan Cheng, "A perceptual two-pass VBR MPEG-2 video encoder", IEEE Transactions on Consumer Electronics, Volume 51, Issue 4, Pages 1237-1247, November 2005.

[466] Yue Yu, Jian Zhou, Yiliang Wang, and Chang Wen Chen, "A novel two-pass VBR coding algorithm for fixed-size storage application", IEEE Transactions on Circuits and Systems for Video Technology, Volume 11, Issue 3, Pages 345–356, March 2001.

[467] Luís Teixeira, and Luís Corte-Real, "Statistical Multiplexing of H.264 Video Streams using Structural Similarity Information", Journal of Information Science and Engineering, Volume 25, Issue 3, Pages 703-715, May 2009.

[468] Aggelos Lazaris, and Polychronis Koutsakis, "Modeling multiplexed traffic from H.264/AVC videoconference streams", Computer Communication, Volume 33, Issue 10, Pages 1235-1242, June 2010.

[469] Aggelos Lazaris, and Polychronis Koutsakis, "Modeling Video Traffic from Multiplexed H.264 Videoconference Streams", Proceedings of the 2008 IEEE Global Telecommunication Conference (GLOBECOM 2008), Pages 1479-1484, 2008.

[470] Mehdi Rezaei, Imed Bouazizi, and Moncef Gabbouj, "Implementing Statistical Multiplexing in DVB-H", International Journal of Digital Multimedia Broadcasting, Volume 2009, Article ID 261231, 2009.

[471] Mehdi Rezaei, Imed Bouazizi, and Moncef Gabbouj, "Fuzzy Joint Encoding and Statistical Multiplexing of Multiple Video Sources with Independent Quality of Services for Streaming over DVB-H", International Journal of Innovative Computing, Information and Control, Volume 5, Issue 6, June 2009.

[472] Mehdi Rezaei, Imed Bouazizi, and Moncef Gabbouj, "Joint Video Coding and Statistical Multiplexing for Broadcasting over DVB-H Channels", IEEE Transactions on Multimedia, Volume 10, Issue 8, Pages 1455-1464, December 2008.

[473] 3GPP R2-074339, E-MBMS functions of statistical multiplexing, RAN WG2 #59bis, Shanghai, China, October 2007.

[474] Cheng-Hsin Hsu, and Mohamed Hefeeda, "Statistical Multiplexing of Variable-Bit-Rate Videos Streamed to Mobile Devices", ACM Transactions on Multimedia Computing, Communications, and Applications, Volume 7, Issue 2, Article 12, Pages 1-23, February 2011.

[475] Farid Molazem Tabrizi, Cheng-Hsin Hsu, Mohamed Hefeeda, and Joseph G. Peters, "Optimal Scalable Video Multiplexing in Mobile Broadcast Networks", Proceedings of the 3rd Workshop on Mobile Video Delivery (MoViD 2010), Firenze, Italy, Pages 9-14, October 2010.

[476] Mohamed Hefeeda, and Cheng-Hsin Hsu, "On Burst Transmission Scheduling in Mobile TV Broadcast Networks", IEEE/ACM Transactions on Networking, Volume 18, Issue 2, Pages 610-623, April 2010.

[477] Cheng-Hsin Hsu, and Mohamed Hefeeda, "On Statistical Multiplexing of Variable-Bit-Rate Video Streams in Mobile Systems", Proceedings of the 17th ACM International Conference on Multimedia (MM 2009), Beijing, China, Pages 411-420, October 2009.

[478] Martin Fleury, Emmanuel Jammeh, Rouzbeh Razavi, Sandro Moiron, and Mohammed Ghanbari, "Video Streaming in Evolving Networks under Fuzzy Logic Control", Christos J Bouras (Editor), "Trends in Telecommunications Technologies", InTech, ISBN 9789533070728, March 2010.

[479] Hamed Ahmadi Aliabad, Sandro Moiron, Martin Fleury, and Mohammed Ghanbari, "No-reference H.264/AVC Statistical Multiplexing for DVB-RCS", Kandeepan Sithamparanathan, Mario Marchese, Marina Ruggieri, and Igor Bisio (Editors), "Personal Satellite Services: Second International ICST Conference, PSATS 2010, Rome, Italy, February 2010 Revised Selected Papers", Springer-Verlag Berlin Heidelberg, ISBN 9783642136177, Pages 163-178, February 2010.

[480] Mehdi Rezaei, Miska M. Hannuksela, and Moncef Gabbouj, "Semi-Fuzzy Rate Controller for Variable Bit Rate Video", IEEE Transactions on Circuits and Systems for Video Technology, Volume 18, Issue 5, Pages 633 - 645, May 2008.

[481] Mehdi Rezaei, Miska M. Hannuksela, and Moncef Gabbouj, "Tune-in Time Reduction in Video Streaming over DVB-H", IEEE Transactions on Broadcasting, Special Issue on "Mobile Multimedia Broadcasting", Volume 53, Issue 1, Part 2, Pages 320 - 328, March 2007.

[482] Mahesh Balakrishnan, Robert Cohen, Etienne Fert, and Gertjan Keesman, "Benefits of statistical multiplexing in Multi-program broadcasting", Proceedings of the International Broadcasting Convention 1997 (IBC 1997), Pages 560-565, September 1997.

[483] Wei-Cheng Gu, and David W. Lin, "Joint Rate-Distortion Coding of Multiple Videos", IEEE Transactions on Consumer Electronics, Volume 45, Issue 1, Pages 159-164, February 1999.

[484] Soon-kak Kwon, K. R. Rao, Oh-Jun Kwon, and Tai-Suk Kim, "Joint Bandwidth Allocation for User-Required Picture Quality Ratio Among Multiple Video Sources", IEEE Transactions on Broadcasting, Volume 51, Issue 3, Pages 287- 295, September 2005.

[485] Mayank Tiwari, Theodore Groves, and Pamela C. Cosman, "Delay constrained multiplexing of video streams using dual-frame video coding", IEEE Transactions on Image Processing, Volume 19, Issue 4, Pages 1022-1035, April 2010.

[486] Mayank Tiwari, Theodore Groves, and Pamela C. Cosman, "Multiplexing video streams using dual frame video coding", Proceedings of the 2008 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2008), Pages 693-696, March 2008.

[487] Mayank Tiwari, Theodore Groves, and Pamela C. Cosman, "Competitive Equilibrium Bitrate Allocation for Multiple Video Streams", IEEE Transactions on Image Processing, Volume 19, Issue 4, Pages 1009-1021, April 2010.

[488] Mayank Tiwari, Theodore Groves, and Pamela C. Cosman, "Bitrate allocation for multiple video streams at competitive equilibria", Proceedings of the 42nd Asilomar Conference on Signals, Systems and Computers, Pages 2248-2252, October 2008.

[489] Mayank Tiwari, Theodore Groves, and Pamela C. Cosman, "Bit-rate allocation for multiple video streams using a pricing-based mechanism", IEEE Transactions on Image Processing, Volume 20, Issue 11, Pages 3219-3230, November 2011.

[490] Mayank Tiwari, Theodore Groves, and Pamela C. Cosman, "Pricing-based decentralized rate allocation for multiple video streams", Proceedings of the 2009 IEEE International Conference on Image Processing (ICIP 2009), Pages 3065-3068, November 2009.

[491] Nesrine Changuel, Bessem Sayadi, and Michel Kieffer, "Predictive encoder and buffer control for statistical multiplexing of multimedia content", IEEE Transaction on Broadcasting, Volume 58, Issue 3, Pages 401-416, September 2012.

[492] Nesrine Changuel, Bessem Sayadi, and Michel Kieffer, "Delay-Sensitive Statistical Multiplexing of Multimedia Contents With Time-Varying Channel Conditions", Proceedings of the 2011 World Wireless Research Forum (WWRF 2011), Qatar, April 2011.

[493] Nesrine Changuel, Bessem Sayadi, and Michel Kieffer, "Statistical multiplexing of distributed video streams", Proceedings of the 2011 GRETSI Symposium on Signal and Image Processing (GRETSI 2011), Bordeaux, France, Pages 1-4, September 2011.

[494] Nesrine Changuel, Bessem Sayadi, and Michel Kieffer, "Joint Encoder and Buffer Control for Statistical Multiplexing of Multimedia Contents", Proceedings of the 2010 IEEE Global Telecommunications Conference (GLOBECOM 2010), Miami, U.S.A., Pages 1-6, December 2010.

[495] Nesrine Changuel, Bessem Sayadi, and Michel Kieffer, "Statistical Multiplexing of Video Programs", IEEE Vehicular Technology Magazine, Volume 4, Issue 3, Pages 62-68, September 2009.

[496] Nesrine Changuel, Bessem Sayadi, and Michel Kieffer, "Rate and Distortion Model for Efficient Statistical Multiplexing of Digital Video Programs", Proceedings of the 2009 World Wireless Research Forum, Paris, May 2009.

[497] Zhihai He, and Dapeng Wu, "Look-ahead processing and statistical multiplexing for multiple JVT video encoders", Proceedings of the 2004 International Packet Video Workshop (PV 2004), August 2004.

[498] Marco Tagliasacchi, Giuseppe Valenzise, and Stefano Tubaro, "Minimum variance optimal rate allocation for multiplexed H.264/AVC bitstreams", IEEE Transactions on Image Processing, Volume 17, Issue 7, Pages 1129-1143, July 2008.

[499] Giuseppe Valenzise, Marco Tagliasacchi, and Stefano Tubaro, "Minimum variance multiplexing of multimedia objects", Proceedings of the 2008 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP 2008), Las Vegas, Pages 1133-1136, April 2008.

[500] Giuseppe Valenzise, Marco Tagliasacchi, and Stefano Tubaro, "A smoothed, minimum distortion-variance rate control algorithm for multiplexed video sequences", Proceedings of the International Workshop on Mobile Video 2007 (MM 2007), Augsburg, Germany, Pages 55-60, September 2007.

[501] Giuseppe Valenzise, Marco Tagliasacchi, Stefano Tubaro, and L. Piccarreta, "A rho-domain rate controller for multiplexed video sequences", Proceedings of the 26th Picture Coding Symposium (PCS 2007), Lisbon, Portugal, Pages 1-4, November 2007.

[502] Paulo Nunes, Grzegorz Pastuszak, Andrzej Pietrasiewicz, and Fernando Pereira, "Joint bit allocation for multi-sequence H.264/AVC video coding rate control", Proceedings of the 26th Picture Coding Symposium 2007 (PCS 2007), Lisbon, Portugal, November 2007.

[503] Andrzej Pietrasiewicz, and Grzegorz Pastuszak, "Rate control for multisequence H.264/AVC compression", Proceedings of the Second International Conference on Signal Processing and Multimedia Applications (SIGMAP 2007), Barcelona, Spain, Pages 456-461, July 2007.

[504] Luís Teixeira, and Teresa Andrade, "Exploiting Characteristics of a large number of MPEG video sources for statistically multiplexing video for TV broadcast applications", Proceedings of the 4th Workshop on Intelligent Methods in Signal Processing and Communications, Bayona, Spain, Pages 65-69, June 1996.

[505] Luís Teixeira, Teresa Andrade, and Vitor Teixeira, "Joint control of MPEG VBR video over ATM networks", Proceedings of the 2007 International Conference on Image Processing (ICIP 1997), Santa Barbara, U.S.A., Volume 3, Pages 586-589, October 1997.

[506] G. Cermak, M. Pinson, and S. Wolf, "The Relationship Among Video Quality, Screen Resolution, and Bit Rate", IEEE Transactions on Broadcasting, Volume 57, Issue 2, Pages 258-262, June 2011.

[507] Quan Huynh-Thua, Matthew Brotherton, David Hands, Kjell Brunnstrom, and Mohammed Ghanbari, "Examination of the SAMVIQ Subjective Assessment Methodology", Proceedings of the Third International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM 2007), Scottsdale, U.S.A., January 2007.

[508] Stéphane Péchard, Romuald Pépion, and Patrick Le Callet, "Suitable methodology in subjective video quality assessment: A resolution dependent paradigm", Proceedings of the Third International Workshop on Image Media Quality and its Applications (IMAQ 2008), Kyoto, Japan, September 2008.

[509] David M. Rouse, Romuald Pépion, Patrick Le Callet, and Sheila S. Hemami, "Tradeoffs in subjective testing methods for image and video quality assessment", Proceedings of SPIE Volume 7527: Human Vision and Electronic Imaging XV, Pages 75270F-75270F-11, 2010.

[510] Rec. ITU-R BT.1788, "Methodology for the subjective assessment of video quality in multimedia applications", Recommendations of the ITU, Radio-communication Sector, 2007.

[511] VQEG Multimedia (MM) Phase I Group, "VQEG MM Testplan (Version 1.21)", March 2008. Available from: http://www.its.bldrdoc.gov/vqeg/projects/multimedia-phase-i/multimedia-phase-i.aspx. Retrieved 31 January 2012.

[512] Wen-Nung Lie, Chih-Fan Chen, and Tom C.-I. Lin, "Two-pass rate-distortion optimized rate control technique for H.264/AVC video", Proceedings of SPIE Volume 5960: Visual Communication and Image Processing 2005, Pages 1061–1070, 2005.

[513] Do-Kyoung Kwon, Mei-Yin Shen, and Chung Chieh Jay Kuo, “R-D optimized frame-layer bit allocation for H.264”, Proceedings of the 2006 International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP 2006), Pages 507–510, 2006.

[514] Chengsheng Que, Guobin Chen, and Jilin Liu, “An efficient two-pass VBR encoding algorithm for H.264”, Proceedings of the 2006 International Conference on Communications, Circuits and Systems (ICCCAS 2006), Guilin, China, Volume 1, Pages 118–122, June 2006.

[515] Jianfei Huang, Jun Sun, and Wen Gao, “A novel two-pass VBR coding algorithm for the H.264/AVC video coder based on a new analytical R-D model”, Proceedings of the 26th Picture Coding Symposium 2007 (PCS 2007), Lisbon, Portugal, November 2007.

[516] Jing Zhang, Thinh M. Le, Sim Heng Ong, and Truong Q. Nguyen, “No-reference image quality assessment using structural activity”, Signal Processing, Volume 91, Issue 11, Pages 2575-2588, November 2011.

[517] Tomás Brandão, and Maria Paula Queluz, “No-Reference Quality Assessment of H.264/AVC Encoded Video”, IEEE Transaction on Circuits and Systems for Video Technology, Volume 20, Issue 11, Pages 1437-1447, November 2010.