ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC

THESIS PRESENTED TO
ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

IN PARTIAL FULFILLEMENT OF THE REQUIREMENTS FOR
THE MASTER DEGREE IN
INFORMATION TECHNOLOGY ENGINEERING
M. Eng.

BY
Alexandru COTOROS PETRULIAN

MACROBLOCK LEVEL RATE AND DISTORTION ESTIMATION APPLIED TO THE COMPUTATION OF THE LAGRANGE MULTIPLIER IN H.264 COMPRESSION

MONTREAL, MARCH 17, 2014

© Copyright 2013 reserved by Alexandru Cotoros Petrulian

© Copyright reserved
It is forbidden to reproduce, save or share the content of this document either in whole or in parts. The reader who wishes to print or save this document on any media must first get the permission of the author.

BOARD OF EXAMINERS

THIS THESIS HAS BEEN EVALUATED

BY THE FOLLOWING BOARD OF EXAMINERS

Mr. Stéphane Coulombe, Thesis Supervisor
Département de génie logiciel et des TI at École de technologie supérieure

Mrs. Rita Noumeir, President of the Board of Examiners
Département de génie électrique at École de technologie supérieure

Mr. Christian Desrosiers, Member of the jury
Département de génie logiciel et des TI at École de technologie supérieure

THIS THESIS WAS PRENSENTED AND DEFENDED

IN THE PRESENCE OF A BOARD OF EXAMINERS AND PUBLIC

ON FEBRUARY 13, 2014

AT ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

ACKNOWLEDGMENTS

I would like to thank my advisor, Professor Stéphane Coulombe, for the chance to perform this research at Vantrix Industrial Research Chair in Video Optimization Laboratory, and for the opportunity to learn and work on such a great research project under his invaluable supervision. I would also like to express my deepest appreciation for his mentorship, encouragement and for sharing his great professional experience throughout my Master's program.

I would like to thank him and the company Vantrix, not only for providing the scholarship which allowed me to pursue this research, but also for giving me the opportunity to attend the group meetings and collaborate with engineers and researchers in the field of video processing.

I would like to thank the board of examiners that gave generously of their time for reviewing this thesis and for their participation in my defense examination.

I would also like to thank Ms. Deepa Padmanabhan and Nick Desjardins for the important contribution they brought to the validation of this project.

Special thanks to anyone who has taken time to take a look, make suggestions and proofread this thesis.

ESTIMATION DU DÉBIT ET DE LA DISTORSION AU NIVEAU DU MACROBLOC APPLIQUÉE AU CALCUL DU MULTIPLICATEUR DE LAGRANGE EN COMPRESSION AVEC LA NORME H.264

Alexandru COTOROS PETRULIAN

RÉSUMÉ

La valeur optimale du multiplicateur de Lagrange (λ), un facteur de compromis entre le débit obtenu et la distorsion mesurée lors de la compression d'un signal, est un problème fondamental de la théorie de la débit-distorsion et particulièrement de la compression vidéo.

Le standard H.264 ne spécifie pas comment déterminer la combinaison optimale des valeurs des paramètres de quantification (QP) et des choix de codage (vecteurs de mouvement, choix de mode). Actuellement, le processus d'encodage est encore dépendant de la valeur statique du multiplicateur de Lagrange, dont une dépendance exponentielle du QP est adoptée par la communauté scientifique, mais qui ne peut pas accommoder la diversité des vidéos. La détermination efficace de sa valeur optimale reste encore un défi à relever et un sujet de recherche d'actualité.

Dans la présente recherche, nous proposons un nouvel algorithme qui adapte de façon dynamique le multiplicateur de Lagrange en fonction des caractéristiques de la vidéo d'entrée en utilisant la distribution des résidus transformés au niveau du macrobloc. Le but recherché est d'augmenter la performance de codage de l'espace débit-distorsion.

Nous appliquons plusieurs modèles aux coefficients résiduels transformés (Laplace, Gaussien, densité de probabilité générique) au niveau du macrobloc pour estimer le débit et la distorsion et étudier dans quelle mesure ils correspondent aux vraies valeurs. Nous analysons ensuite les bénéfices et désavantages de quelques modèles simples (Laplace et un mélange de Laplace et Gaussien) du point de vue du gain en compression et de l'amélioration visuelle en rapport avec le code de référence du standard H.264 (amélioration débit-distorsion).

Plutôt que de calculer le multiplicateur de Lagrange basé sur un seul modèle appliqué sur toute la trame, comme proposé dans l'état de l'art, nous le calculons basé sur des modèles appliqués au niveau du macroblock. Le nouvel algorithme estime, à partir de la distribution des résidus transformés du macrobloc, le débit et la distorsion de chacun, pour ensuite combiner la contribution de chacun pour calculer multiplicateur de Lagrange de la trame.

Les expériences sur des types variés de vidéos ont démontré que la distorsion calculée au niveau du macrobloc est proche de la distorsion réelle offerte par le logiciel de compression vidéo de référence pour la plupart des séquences vidéo testées, mais un modèle fiable pour le débit est encore recherché particulièrement à très bas débit. Néanmoins, les résultats de compression de diverses séquences vidéo montrent que la méthode proposée performe beaucoup mieux que le Joint Model du standard H.264 et un peu mieux que l'état de l'art.

MACROBLOCK LEVEL RATE AND DISTORTION ESTIMATION APPLIED TO THE COMPUTATION OF THE LAGRANGE MULTIPLIER IN H.264 COMPRESSION

Alexandru COTOROS PETRULIAN

ABSTRACT

The optimal value of Lagrange multiplier, a trade-off factor between the conveyed rate and distortion measured at the signal reconstruction has been a fundamental problem of rate distortion theory and video compression in particular.

The H.264 standard does not specify how to determine the optimal combination of the quantization parameter (QP) values and encoding choices (motion vectors, mode decision). So far, the encoding process is still subject to the static value of Lagrange multiplier, having an exponential dependence on QP as adopted by the scientific community. However, this static value cannot accommodate the diversity of video sequences. Determining its optimal value is still a challenge for current research.

In this thesis, we propose a novel algorithm that dynamically adapts the Lagrange multiplier to the video input by using the distribution of the transformed residuals at the macroblock level, expected to result in an improved compression performance in the rate-distortion space.

We apply several models to the transformed residuals (Laplace, Gaussian, generic probability density function) at the macroblock level to estimate the rate and distortion, and study how well they fit the actual values. We then analyze the benefits and drawbacks of a few simple models (Laplace and a mixture of Laplace and Gaussian) from the standpoint of acquired compression gain versus visual improvement in connection to the H.264 standard.

Rather than computing the Lagrange multiplier based on a model applied to the whole frame, as proposed in the state-of-the-art, we compute it based on models applied at the macroblock

level. The new algorithm estimates, from the macroblock's transformed residuals, its rate and distortion and then combines the contribution of each to compute the frame's Lagrange multiplier.

The experiments on various types of videos showed that the distortion calculated at the macroblock level approaches the real one delivered by the reference software for most sequences tested, although a reliable rate model is still lacking especially at low bit rate. Nevertheless, the results obtained from compressing various video sequences show that the proposed method performs significantly better than the H.264 Joint Model and is slightly better than state-of-the-art methods.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

LIST OF ABBREVIATIONS

AC	Non-zero frequency (of transformed coefficients)
AVC	Advanced Video Coding
CABAC	Context-based Adaptive Binary Arithmetic Coding
CAVLC	Context-based Adaptive Variable-Length Coding
CIF	Common Intermediate Format
DC	Zero frequency (of transformed coefficients)
DCT	forward Discrete Cosine Transform
DPB	Decoded Picture Buffer
EPZS	Enhanced Predictive Zonal Search
FPel	Motion compensation by interpolation with integer-pixel (Full Pel) accuracy
GGD	Generalized Gauss Distribution
HPel	Motion compensation using interpolation with half-pixel accuracy
HVS	Human Visual System
i.i.d.	independent and identically distributed (random variable, process)
IDCT	Inverse Discrete Cosine Transform
JM	Joint Model (reference software)
MB	Macroblock
MC	Motion Compensation
MCU	Multiple Coding Units
MD	Mode Decision
ME	Motion Estimation
MOS	Mean Opinion Score
MSE	Mean Squared Error
MV	Motion Vector
pdf	probability density function
QCIF	Quarter Common Intermediate Format
QP	Quantization Parameter
QPel	Motion compensation using interpolation with quarter-pixel accuracy
RD	Rate-Distortion

RDO	Rate-Distortion Optimization
SAD	Sum of Absolute Differences
SSE	Sum of Squared Errors
SSIM	Structural Similarity Index
VLC	Variable-Length Code

LIST OF SYMBOLS

λ_{mcu}	Lagrange multiplier using the MCU approach
λ_{HR}	Lagrange multiplier determined with high rate approximation
λ_{MODE}	Lagrange multiplier for the mode decision stage
λ_{MOTION}	Lagrange multiplier for the motion estimation stage
λ	Lagrange multiplier
Λ	Laplace parameter
η	location (expected value of Laplace distribution)
θ	scale
μ	mean value (expected value of Gauss distribution)
σ	standard deviation
γ	rounding offset of the quantization step
α	shape parameter describing the exponential rate of decay (GGD)
β	standard deviation (GGD)
D	distortion
J	cost function
Q	quantizer step size
R	bit rate
H	entropy

INTRODUCTION

Context

This research endeavors to find the highest quality of the video that would allow exporting a reasonably smaller amount of bits, so that the overall coding gain is superior to the actual one.

Problem statement

The fundamental problem in video compression is to obtain the best trade-off between the conveyed rate and the perceived distortion of the reconstructed video. The H.264 standard has only the syntax standardized, but it does not specify how the optimal values, e.g. the Lagrange multiplier, may be obtained, nor the best encoder configuration or coding decisions. Rate-distortion optimization is originally posed as a constrained problem of finding the minimum distortion under a rate constraint. For simplicity, this problem is converted into an unconstrained optimization one using a Lagrangian method in which the Lagrange multiplier needs to be carefully selected. The rate, as a byproduct of distortion calculation and Lagrange multiplier determination, is not always well-calculated and it is a challenge to find its best value in combination with the other two. On the other hand, the Lagrange multiplier depends on the video source features, while the distortion should be calculated with the help of a modern visual quality metric, based on the human visual system (HVS). Despite these difficulties, it is believed that there is enough room for significant improvement of the bit allocation with video quality improvement, for any type of video content. It is supported by the huge amount of research in this direction.

In the end, the problem of finding the optimal value of Lagrange multiplier that would determine balanced values of the pair (distortion, rate) from the standpoint of video quality and bit allocation is fundamental in these respects. Huge investments have been injected in the related industries (entertainment, communication, social media, business, defense, health,

video surveillance, traffic management) that made video so widespread and continually growing expressing the need to visually communicate at higher resolutions. YouTube, the largest and most popular user-generated video-hosting site, has demonstrated the insatiable and widespread demand for video content. In 2011, the site reported one trillion video viewings in total - an average of 140 viewings per capita, on the entire world population; and reports that approximately four billion hours of video are watched per month (Atkinson, 2012). These staggering figures apply for just one site - and the demand is ever-growing. Indeed, many broadcasting corporations have responded to the increasingly web-based desire for content and have made television shows available for streaming off their individual online sites (Barker, 2011). Ever since the slow but sure decline of video rental businesses like Blockbuster (Carr, 2010), which loaned popular titles on physical disks, other companies have taken up the torch of modern video subscription services. Netflix has seen its consumer base grow to 30 million subscribers in 2012 (Etherington, 2012), and other providers such as iTunes (Yarow, 2013) and Amazon Video (Stone, 2013) have been growing steadily as well.

One of the most important factors in the exponential growth of video demand is the combination of social integration with increasingly more popular mobile devices. YouTube reports 500 years of video being watched every day just through links on social platforms like Facebook, and 700 videos being shared every minute on Twitter (Atkinson, 2012). Bandwidth demand for mobile video has exploded since the growth of ever-more capable portable devices, especially in the smartphone and tablet sector (Kovach, 2013). The adoption of and desire for high-quality video has been a leading force in the hunger for greater bandwidth capacity, speed and reliability, in the form of residential, corporate and cellular data availability. Especially in the case of the latter, telecommunications companies have invested billions of dollars into emerging technologies, like long-term evolution networks that ensure faster and more reliable data transfers, while increasing bandwidth capacity. The mass-adoption of these technologies by consumers will ensure their long-term sustainability, although predictably, speeds are bound to suffer as more traffic is introduced. The heaviest burden on these networks is undoubtedly the transfer of video content – whether it is video conferencing over services like Microsoft's Skype or Apple's Face Time, watching

television (TV) shows on-the-go, or uploading home videos straight from mobile devices. Clearly, there is an increasing demand to communicate and share information today that is only the beginning of the years to come.

Short literature overview

The Lagrange multiplier λ, the balancing factor between rate and distortion in rate-distortion theory, evolved with the versions of the video standards. While H.263 used an expression depending on the square of the quantization parameter (QP) only, H.264/AVC (Advanced Video Coding) promoted a value that grows exponentially with QP, but both expressions of λ are static, regardless of the sequence. The state-of-the-art approach (Li and others, 2009) has set the trend of using variable Lagrange multipliers, adaptive with the statistical properties of the video content, which resulted in a different value of λ for each frame. This method allows getting compression improvements, especially on slow paced videos. New visual quality metrics closer to HVS, such as the structural similarity index (SSIM) have emerged, replacing the sum of squared errors (SSE) and giving a boost to λ calculated per frame and adjusted at the macroblock level as (Wang and others, 2012) propose.

The thesis is organized as follows: Chapter 1 centers on basic concepts related to video compression and rate-distortion theory whereas chapter 2 presents the state-of-the-art approaches and various types of distributions for the transformed residual coefficients. In chapter 3, the focus is on the Laplace and Gauss equations for rate and distortion, while the fourth one outlines the methods to determine the Lagrange multiplier at the macroblock level. Chapter 5 presents the experimental results. Chapter 6 concludes this thesis.

Contributions

This work proposes a new algorithm to dynamically adapt the Lagrange multipliers based on the distribution of transform residuals at the macroblock level, whose purpose is to improve the performance in terms of rate-distortion.

We study the estimation of rate and distortion functions, at the macroblock level, using various probability distribution functions for the transformed residuals, choosing from Laplace, Gauss, generalized Gauss and a generic, numerically computed, probability density function (pdf). We then conceive an algorithm for computing the Lagrange multiplier based on each macroblock's rate and distortion functions. This permits selecting the most appropriate pdf per macroblock instead of assuming that a single distribution applies to the whole frame. This permits to estimate adaptively the most appropriate λ for each frame based on macroblock's statistics and use it both at the motion compensation and mode decision stages of the compression.

CHAPTER 1

AN OVERVIEW OF VIDEO COMPRESSION

In this chapter, the fundamental notions and basic functionality of the H.264 standard are presented, followed by a description of the Lagrangian multiplier technique, as an efficient way to solve the rate distortion optimization problem. Finally, a block diagram summarizes the way the rate distortion optimization mechanism integrates with the main coding flow and how their functional relationship may be exploited for the benefit of the video compression.

1.1 Basic concepts in the H.264 standard

H.264/MPEG-4 Part 10 AVC (Advanced Video Coding) is the standard for video compression that is the most widespread. It is based on concepts such as prediction, motion estimation, motion compensation, mode decision, transformation, quantization, entropy encoding, deblocking, visual quality, that are described next.

1.1.1 Prediction

The compression performance of the encoder depends on the efficiency of the prediction methods. In order to create a slim residual, as scarce as possible of non-zero data, an accurate prediction is detected and extracted from the original macroblock. Thus, the best match of the current block, chosen inside one of the designated reference frames, is chosen so as to minimize the necessary bits to encode the motion vectors.

The prediction block found is further used to generate the residual transformed coefficients.
In the case of the intra prediction, the best match of an I-type macroblock is searched by using the adjacent and previously coded blocks in the same slice (frame). In this way, one exploits the existing spatial correlation between the current block to be encoded and its

neighbors. The best intra prediction is searched at different block sizes. In AVC there are 9 possible prediction modes for a 4x4 and 8x8 luma blocks as illustrated in Figure 1.1.
For a luma macroblock or chroma block there are four possible intra prediction modes in AVC as illustrated in Figure 1.2. In order to increase the coding efficiency, the most probable prediction mode is calculated as the starting point before entering the search phase.

Since the nearest samples in the signal are not fully independent and identically distributed (i.i.d.), high correlations between them exist in the temporal domain, i.e. between temporally adjacent frames. The correlation degree increases with the sampling rate.

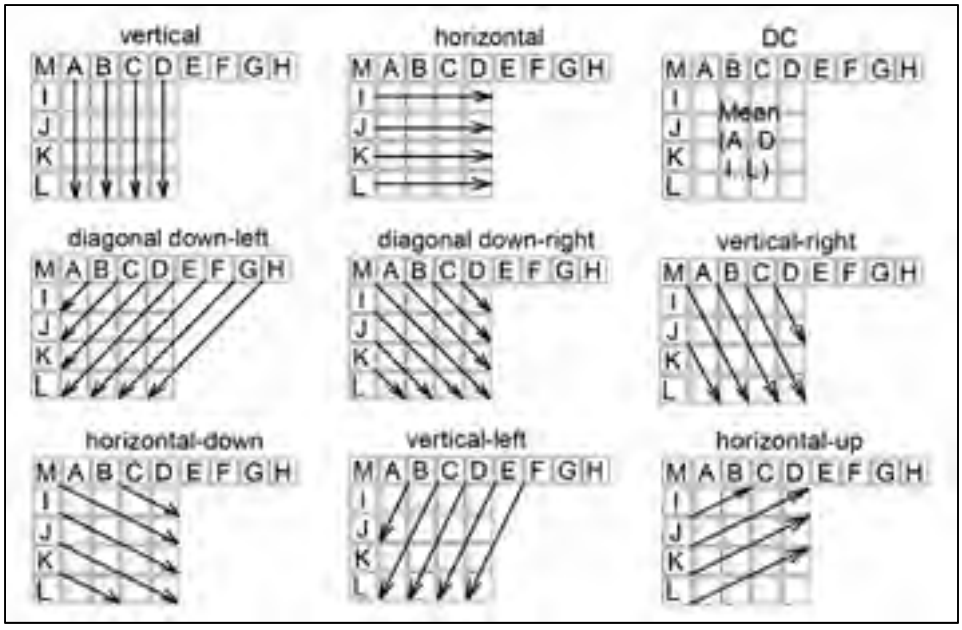


Figure 1.1 4x4 intra prediction modes.
Adapted from (Richardson, 2010)

Inter prediction comes into play, taking advantage of the previously reconstructed frames, available in the decoded picture buffer (DPB), thus motion compensating the encoding with the offset between the original and its prediction. The macroblock/block estimate is searched in a region, usually a 32 pixel square, centered on the original macroblock.

The current macroblock can be predicted as predicted (P) type (also called inter) when the samples chosen as reference are selected from the list of past encoded frames or bidirectional (B) type, in which case, the prediction is based on samples in the list of past and respectively the list of future encoded frames, from the standpoint of displaying order.

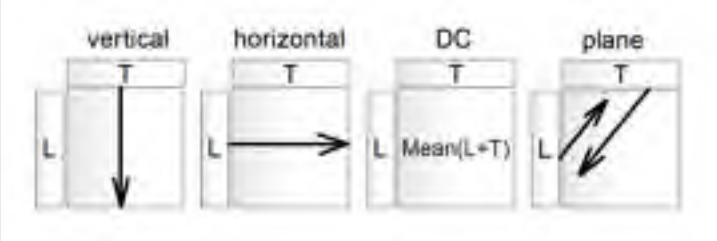


Figure 1.2 Intra 16x16 prediction modes.
Adapted from (Richardson, 2010)

A (P/B) Skip mode, which is only permitted in P/B slices, occurs when no data – MVs (motion vectors) differences and transformed residual coefficients - are transmitted to the decoder. Yet, the macroblock data is reconstructed at the decoder, through interpolation of the previously coded data, using the motion compensated prediction with a MV derived from previously sent vectors of a single reference frame in the case of P-Skip mode, or from two adjacent reference frames in the case of B-Skip Direct mode.

When the motion in the scene is so complex that the macroblock size would be too big to observe it in detail, the macroblock is divided into partitions (8x16, 16x8, and 8x8) and further sub-partitions of the block 8x8 (4x8, 8x4 and 4x4) as illustrated in Figure 1.3. Though the partitioning has the drawback of increasing the amount of motion vectors to transmit, looking for smaller partitions (8x8, 4x8, 8x4, and 4x4) has the benefit of decreasing the energy of the signal difference between the original and the best match. The optimal prediction of a macroblock (block) is always accompanied by its associated motion vector.

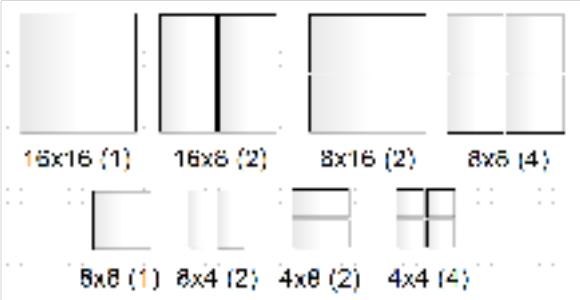


Figure 1.3 Macroblock/sub-macroblock partitions for interframe coding.
Adapted from (Richardson, 2010)

The motion vector refers to the offset between the positions of the current partition and its best prediction in the reference frame(s). It can point to integer, half or quarter pixel positions in the luma component of the reference picture, depending on the pixel accuracy with which the search is performed. The half-pixel luma samples are generated using a 6 tap finite impulse response (FIR) filter applied to integer-pixel samples in the reference frame, while the quarter-pixel samples are inferred through linear interpolation between adjacent half pixel samples. Alternately, in order to increase the motion accuracy at 4:2:0 resolution, the quarter pixel positions are calculated for chroma samples using a 4 tap FIR to interpolate between the neighboring integer and half pixel positions.

The motion estimation process, available for inter-prediction only, defines the space of pair solutions (predicted $region_i$, MV_i) that are searched upon. The best prediction does not necessarily involve the minimum effort to encode its motion vector (MV). Although small (sub) partitions offer the best estimate, encoding their motion vectors can incur a significant number of bits. The spatial and temporal correlation between nearby partitions is often found at the level of their motion vectors; hence the motion vector of a block can be predicted from those associated to the previously coded blocks. The motion vectors prediction applies spatially, by considering the median of the surrounding blocks MVs. What is encoded in the macroblock's header is the difference between the current MV and the predicted motion vector MVp.

The outcome of the prediction phase for the current block is the estimated block and the associated motion vector that is used to motion compensate the encoded macroblock with respect to reference frame(s). Only the motion vectors differences are encoded in the bitstream, namely in the macroblock header, while the prediction is further used to shape the coefficients of the residual block.

Among the most efficient methods of finding the best prediction are the zonal search algorithms represented by PMVFAST (Tourapis, Au and Liou, 2001) and its extension, Enhanced Predictive Zonal Search or EPZS for short (Tourapis, 2002). PMVFAST enhances the speed and video quality by considering the following as initial predictors: the motion vectors of spatially adjacent blocks in the current frame, the (0, 0) motion vector, the median predictor, and the motion vector of the collocated block in the previous frame, all, as a matter of temporal domain correlation. It introduces reliable early-stopping criteria, at any check-point, based on correlations between adjacent blocks, though fixed thresholds are used to compare with the sum of absolute differences (SAD) values.

The highly efficient EPZS algorithm, improves upon PMVFAST by considering at prediction stage several highly likely predictors, based on multi-stage checking pattern. Key to its performance is the fact that the MVs of the current block can be highly correlated with the MVs of the spatially and temporally adjacent blocks and the introduction of the accelerator MV, to model the variable speed movement of the collocated block with respect to the previous two encoded frames. The current block can also be highly correlated with the adjacent blocks to the collocated block in the previous frame. The effect of these last predictors translates into decreasing entropy of the differences between EPZS MVs compared to PMVFAST MVs.

SAD, the distortion measure between the current frame I_t and i-th previously encoded frame I_{t-i}, displaced by MV with the components (v_x, v_y),

$$SAD(v_x, v_y) = \sum_{m,n=1}^{M,N} | I_t(x+m, y+n) - I_{t-i}(x+v_x+m, y+v_y+n) | \qquad (1.1)$$

where $M, N \in \{4,8,16\}$

is compared to the thresholds that are adaptively calculated in the case of EPZS. Simplified search patterns, square or diamond of order one proved beneficial, significantly reducing the number of checking points and algorithm complexity.

The result of EPZS is a significant reduction of bits necessary to encode the MVs.

1.1.2 Residual coefficients

The prediction stage is lossless and partially removes redundancy by extracting the best matching estimation from the current macroblock (partition/sub-partition). The process of motion compensation takes into account the difference between their positions in terms of motion vector differences (MVD). The more precise the prediction is, the less energy in the residual remains, and the data becomes easier to compress to lesser bits. The residuals, as a difference between the original and predicted signals, contain much less energy than either component, so it requires fewer amounts of bits to be sent to the decoder.

As a matter of fact, since the best mode decision depends on finding the most suitable Lagrange multiplier to encode the macroblock, modeling the residuals has become a central problem in rate-distortion optimization (RDO).

Although the majority of natural phenomena and processes statistically behave in the Gaussian way, the residuals in video compression do not quite follow the same path. Similar to audio signals compression, the residuals should behave according to Laplace probability distribution only. In reality, as will be shown in chapter 4, at smaller QP, almost every sequence contains a percentage of Gaussian-distributed residual coefficients, and there is a tendency of macroblock residuals' shape to morph from Laplace to Gauss distribution, when QP decreases.

The central problem of process modeling from the standpoint of macroblock distribution of residuals has been to determine the suitable kind of probability density function (pdf), by adequately evaluating its correctness according to the existing criteria of goodness of fit. Further, using several mathematical models of distortion and entropy, based on distribution of residuals, one can estimate better expressions for Lagrange multipliers.

1.1.3 Transform and quantization

The purpose of DCT is to further de-correlate, compact and translate the residual data into the frequency space, represented by the DC(zero frequency)/AC(non-zero frequency) transformed coefficients. In H.264, non-unitary and signal independent core matrices are defined for the stages of forward and inverse transforms, respectively of a 4x4 block:

$$C_{f4x4} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 2 & 1 & -1 & -2 \\ 1 & -1 & -1 & 1 \\ 1 & -2 & 2 & -1 \end{bmatrix} \quad C_{i4x4} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1/2 & -1/2 & -1 \\ 1 & -1 & -1 & 1 \\ 1/2 & -1 & 1 & -1/2 \end{bmatrix} \tag{1.2}$$

Alternatively, the transforms C_{f8x8} and C_{i8x8} for 8x8 blocks processing exist.

The purpose of the transform is to decorrelate the input signal X from the product $C_{f4}.X.C_{f4}^{T}$. Since the forward transform is not perfectly unitary, a diagonal matrix is hard to obtain, yet, beneficial is the fact that the signal energy is compacted in as little coefficients as possible.

Figure 1.4 The forward transform and quantization.
Adapted from (Richardson, 2010)

At the end of the forward transform illustrated in Figure 1.4, the energy gets redistributed and concentrated into a smaller number of coefficients which makes it easier for entropy encoding. Although the product of orthogonal matrices for forward (C_{f4x4}) and inverse (C_{i4x4}) transform, even normalized, is not perfectly equal to unit matrix, their elements were intentionally multiplied/rounded to get a minimal set $\{\pm 1, \pm 2, \pm 1/2\}$ of values that make the transform easily implementable, by using only additions and left/right bit shifts. This way one can avoid overwhelming floating point multiplication.

Due to the reversibility of the integer DCT, whose inverse matrix can be obtained through the transposing of the normalized core transform (unitary), the overall forward/inverse transform is lossless as well, as is the prediction stage.

In the case of 16x16 intra coded luma blocks and all-dimensions chroma blocks, the DC coefficients are further de-correlated through a DC 4x4 Hadamard transform. In the H.264/AVC standard, the transform and quantization phases overlap in order to minimize the computational effort that would otherwise be overwhelming for the processing unit(s) had they been performed separately.

In the intra-frame coding, DCT is applied to the macroblocks pertaining to the frame itself, while in inter-frame compression its input is defined as the difference between the current block and its prediction.

In addition, a normalization step, necessary to orthonormalize the core integer transform, is integrated with the quantization phase in order to reduce the number of multiplications. Up to this point, taking advantage of the spatial and temporal redundancy and de-correlating the signal, the prediction and transform stages are deemed as lossless steps.

The quantization process is the only lossy phase in the encoder, accounting for the trade-off between the compression performance and the perceived visual quality.

Figure 1.5 Inverse quantization and transform.
Adapted from (Richardson, 2010)

At the end of the re-scaling and inverse transform illustrated in Figure 1.5 the reconstructed macroblock that emerges can be compared against the original one in order to assess the distortion.

While from the energy compaction standpoint, the Karhunen–Loève transform (KLT) is the best method. Its transformation matrix depends on the input signal statistics and lacks computational speed compared with the discrete cosine transform (DCT).

The quantization seen as a down scaling/re-scaling process of signal discretization /reconstruction is built on the linear scalar scheme containing the dead zone (DZ), an uniform threshold scalar quantization (UTSQ), and a nearly uniform reconstruction quantization (NURQ) as described in (Sun and others, 2013a) and (Wang, Yu and He, 2011). The rounding offsets (*z, f*) domain of the forward quantization and reconstruction, illustrated in Figure 1.6, are determined as:

$$z \in (0.5\ldots1),\ f \in (0\ldots0.5) \text{ under the linear constraint } z - f = c,\ c \in (0.5\ldots1) \qquad (1.3)$$

with optimal values for intra (z = 2/3) and inter (z =5/6) coding.

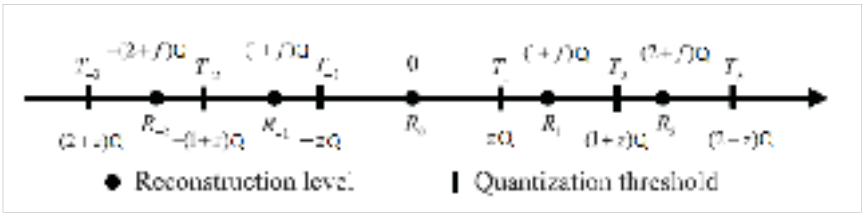


Figure 1.6 DZ + UTSQ/NURQ scheme.
Adapted from (Sun and others, 2013a)

A non-linear quantizer would appropriately reduce the amplitude of the transformed data, but would have to adaptively configure the threshold parameters (*z, f*), based on the density of the transformed signal's pdf.

1.1.4 Mode decision and the macroblock encoding

Mode selection is the process of determining the best block partitions to encode a macroblock. It is governed by the RDO process, as presented in pseudo code in (Richardson, 2010) which takes into account the available modes shown in Figure 1.7.

For every macroblock
  For every available coding mode *m*
    Code the macroblock residual through DCT and quantization using the specific MVs for that mode *m*
    Calculate *R*, the number of bits required to code the macroblock
    Reconstruct the macroblock through inverse quantization and IDCT
    Calculate *D*, the distortion between the original and decoded macroblock
    Calculate the mode cost J_{MODE} , with appropriate choice of Lagrange multiplier
  end
  Choose the mode that gives the minimum J_{MODE}
End

The calculation of the mode cost J_{MODE} was presented in section 1.2.1.

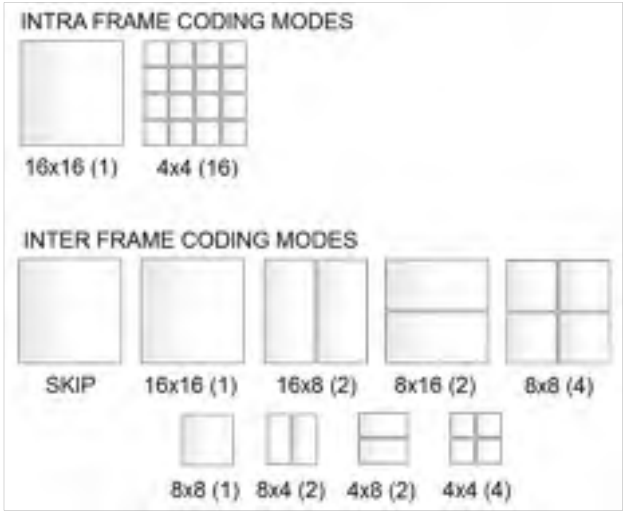


Figure 1.7 Available prediction modes.
Adapted from (Richardson, 2010)

1.1.5 The bit cost of coding a macroblock

During the mode decision stage, when various modes are tested, the final decision is made when the minimum is attained by the cost function J_{MODE}. The number of bits necessary to encode with the best mode is given in principal by two components (Richardson, 2010):

- The header bits, which contain the macroblock mode (I/intra, P/B inter coded), the prediction parameters (the entry (es) in the reference list(s), motion vectors differences for P/B macroblock, the search accuracy: full pel, (FPel), half pel (HPel), quarter pel (QPel).

- The transform coefficients bits as the bits necessary to encode the quantized transformed coefficients, CBP, QP, optimal mode.

1.1.6 Entropy encoding

The entropy coding is the last lossless stage of reducing the video information redundancy. The quantized coefficients are reordered through a zigzag scan to group together the non-zero (DC) values at the beginning of the run-length encoding, followed by the higher frequencies (AC) coefficients, most of them being runs of zeroes.

CAVLC (Context-based Adaptive Variable-Length Coding) maps the coefficients to a series of variable length codewords, using Huffman codes, where frequently-occurring symbols are represented with short variable-length code (VLC) (Richardson, 2010). It uses a context adaptive scheme based on several VLC look-up tables containing the updated statistics of the symbols to encode.

Unlike CAVLC, whose drawback is the assignment of an integral number of bits for each symbol, CABAC (Context-based Adaptive Binary Arithmetic Coding) encodes the whole message by mapping it to a subunit number. It uses context models (probability tables) and binarization schemes that feed the arithmetic coding engine with the necessary updated statistics of the symbols thus eliminating the multiplications operations. CABAC, though

slower than CAVLC, achieves a better compression by allowing fractional number of bits to represent a symbol, thus approaching the theoretical optimal compression ratio.

1.1.7 Visual quality and encoding performance indexes

Of all well-known measures utilized to compute the distortion between the original *I* and the decoded *I'* images with resolution of (*M.N*) pixels,

$$D_{SSE} = \sum_{x,y=1}^{M,N} [I(x,y) - I'(x,y)]^2 \tag{1.4}$$

$$D_{MSE} = D_{SSE} / (M.N)$$

$$D_{SAD} = \sum_{x,y=1}^{M,N} | I(x,y) - I'(x,y) | \tag{1.5}$$

$$D_{SATD} = \alpha \sum_{x,y=1}^{M,N} | T(I(x,y) - I'(x,y)) | \text{, where T = Hadamard transform} \tag{1.6}$$

the metric MSE (mean squared error) is given preference for its meaning – the energy of the error signal (Wang and Bovik, 2006) and because it is preserved through unitary transform. Despite being deemed as an objective visual quality measure, it is poorly correlated with the perceived image quality. Peak signal-to-noise ratio (PSNR), as an objective visual performance index, which is based on MSE, it does not relate with the human perception as well.

$$PSNR = 10 \log_{10} \left(\frac{255^2}{D_{MSE}} \right) \tag{1.7}$$

Nevertheless, Table 1.1., as outlined in (Bouras and others, 2009) summarizes the correlation between the values of PSNR and the perceived visual quality levels stated by the mean opinion score (MOS), as a subjective visual quality index. Through its mapping to MOS, PSNR gets an additional feature that brings it closer to the human visual perception.

Table 1.1 PSNR to MOS mapping.
Adapted from (Bouras and others, 2009)

PSNR(dB)	MOS
>37	Excellent (5)
31-37	Good (4)
25-31	Fair (3)
20 25	Poor (2)
<20	Bad (1)

For this reason and the fact that PSNR is still the performance index of choice, PSNR was adopted as the index to measure the visual quality of our experiments.

1.2 Rate distortion optimization in H.264

Since the video sequences mainly contain motion (quantified as motion vectors) and content (coefficients resulting from techniques to reduce spatial and temporal redundancy, quantified as luma and chroma total runs and trailing ones), the task of the encoder is to find the optimal set *I* of options of the coding parameters, i.e. encoding mode and side information (MVs-motion vectors, macroblock type, skip information, delta QP), so that the distortion is minimized and the resulting bitstream does not surpass a maximum allowable bandwidth.

The central problem of the rate distortion optimization consists in solving the bit allocation approach, which has the constrained form as described in the equation (1.1).

$$\min_{I} D(S,I) \text{, where } R(S,I) \le R_C \tag{1.8}$$

The terms *D(S, I)*, whose minimum is looked for, and *R(S, I)* represent the total additive distortion and rate respectively, for a quantized source signal *S* under an optimal set *I* of options, chosen during the encoding. *I* may include an efficient motion estimation (ME) method like *Enhanced Predictive Zonal Search* (EPZS), appropriate QP range, decision thresholds, reconstructed levels, rounding offsets.

1.2.1 Cost function

A practical, unconstrained form, useful to the discrete codec, looks to achieve the global minimum of the cost function *J*:

$$J(S,I \mid \lambda,Q) = D(S,I,Q) + \lambda.R(S,I,Q) \tag{1.9}$$

for a certain value of the Lagrangian parameter λ that multiplies the rate term and it is referred to as the bit-allocation technique using the Lagrange multiplier (Wiegand and others, 2003). In the equation above, Q is the quantizer value which is related to the quantization factor QF.

The value of Lagrange multiplier can be determined for the convex hull of the rate-distortion (RD) curves as:

$$\frac{\partial J(S,I \mid \lambda,Q)}{\partial R} = 0 \tag{1.10}$$

Thus

$$\lambda = -\frac{\partial D}{\partial R} = -\frac{\partial D/\partial Q}{\partial R/\partial Q} \tag{1.11}$$

Finding its optimal value is a challenge for the research. With the new formulation, it is not anymore necessary to look for the minimum value of the distortion, since a zero distortion would lead to a large bitrate. Instead, a trade-off between the distortion and rate, attainable through a certain set *I* of coding parameters, can lead to an overall global minimum of *J*. The sum of the minimums of the local J_{MB} cost functions, calculated for each macroblock MB with the optimal coding options for ME and MD (mode decision), would result in the minimum of *J* function at the frame level if we make the assumption that coding of MBs are independent. The coding options may contain, among other parameters, the frame type (INTER, INTRA), the transform coefficients values, the quantizer value Q, the motion vectors for interframe, the reference list index(es) pointing to previously encoded frames. Each macroblock has multiple mutual temporal and spatial dependencies with the neighbors in the same frame or former/next encoded frames, which induces a large dependency and

complexity in the codec, making from it a NP-hard problem, thus preventing from deriving a tangible analytical form of global *J*. The difficulty to solve the optimal codec increases with the frame's resolution. The Cartesian product of all coding mode parameters forms the space from which the optimal combination is selected that minimizes the cost function. For progressive-scanned video H.264/AVC seven possible macroblock modes (INTRA4x4, INTRA16x16, SKIP, INTER16x16, INTER16x8, INTER8x16 and INTER8x8) are considered along with 3 sub-macroblock types (INTER8x4, INTER4x8, INTER4x4) available in INTER8x8 mode only. The optimal bit allocation method using Lagrangian multiplier in inter modes encoding applies to both to ME and MD, in this order.

The Lagrange multipliers method is firstly applied at the ME level to find the best match in the decoded reference frame(s). Unlike the MD stage, the ME optimization process calculates the motion compensation distortion between the original and matching block, displaced with using the motion vector (MV), while the rate refers to the bits to encode MVs difference.

The variable number of modes used to find the best macroblock match is based on a similar cost function minimization, which depends on the search method and the refinement degree (FPel, HPel, or QPel). PMVFAST and EPZS have the best results in terms of search time. The successful MV candidate $m_s(MB)$ for the macroblock MB is found by solving the equation:

$$m_s(MB) = \underbrace{\arg\min J_{MOTION}(MB, m_i)}_{m_i \in \{MV\}} \quad \text{, where} \tag{1.12}$$

$$J_{MOTION}(MB, m_i) = D_{DFD}(MB, m_i) + \lambda_{MOTION} . R_{MOTION}(MB, m_i)$$

where $D_{DFD}(MB, m_i)$ is the distortion of the displaced frame difference calculated between the original macroblock and the predicted one, displaced with the motion vector m_i, according to (Wiegand and Girod, 2001). The term $R_{MOTION}(MB, m_i)$ stands for the bitrate necessary to encode each separate m_i candidate. The phase of finding the MV is performed for interframe coding only.

The outcome of the ME stage is represented by the signal difference *S(MV)* and the allotted bitrate to encode the final MVs, as displacements of all blocks, relative to the reference frames. The signal *S(MV)* that contains the prediction error between the predicted and original MB, is further processed - transformed, quantized, entropy coded - during the MD phase, to get the rate (dependent on the transformed coefficients, quantization step, side information and implicitly MV).

For the MD phase the optimization problem becomes:

$$I^* = \underbrace{\arg\min J_{MODE}(S_k, I_k \mid Q)}_{I_k \in \{\text{modes}\}} \quad \text{, where} \tag{1.13}$$

$$J_{MODE}(S_k, I_k \mid Q) = D_{REC}(S_k, I_k \mid Q) + \lambda_{MODE}.R_{REC}(S_k, I_k \mid Q)$$

where the S_k denotes the macroblock partition given the coding option I_k and quantizer step Q, D_{REC} represents the distortion between the original and reconstructed MB for the coding option I_k, while R_{REC} stands for the rate obtained through entropy encoding. The bit allocation in the process of finding the best mode is performed for both intraframe/interframe coding and may include the SKIP mode.

All the aforementioned macroblock and sub-macroblock modes I_k are tested; the one (I^*) whose cost function value is minimal is selected. Thus, the role of the mode decision is to further refine the signal previously acquired during the motion compensation coding and find the best rate by testing for optimal set of encoding parameters.

1.2.2 Optimal Lagrange multiplier

The optimal value of the Lagrangian parameter depends on multiple interdependent parameters; its value must be adjusted in accordance to their values, at frame level and further refined at MB (block) level. Among those parameters, the quantization step Q, the side information (header bits, MV, quantized zeroes), the DCT (forward Discrete Cosine Transform) transformed and quantized coefficients are the most important.

Wiegand and Girod (Wiegand and Girod, 2001) have investigated the relationship between an efficient λ and the DCT plus the scalar quantizer. The expression of λ_{MODE} was experimentally deduced upon encoding the INTER frames of several test sequences with various values of λ_{MOTION}, QP and distortion metrics (SAD, SSE).

$$\lambda_{MODE \mid SAD \,\&\, SSE} = (0.85).2^{\frac{QP-12}{3}} \qquad (1.14)$$
$$\lambda_{MOTION \mid SAD} = \sqrt{\lambda_{MODE}}$$
$$\lambda_{MOTION \mid SSE} = \lambda_{MODE}$$

Evidently, the only parameter of the static λ is the quantization parameter QP, as opposed to the Q step that is generally used when determining λ based on the statistical features of the input sequence as in (Li and others, 2009) and (Wang and others, 2012). While QP ranges from 1 to 51, Q step encompasses a domain 0.625...~230. For the same interval, λ_{MODE} values are (0.067… 6963), while λ_{MOTION} values belong to the interval (0.26…83.44) when the distortion metric is SAD.

The Lagrange multiplier determined with the formulae above is static and does not depend in any way on the sequences' characteristics (the type of distribution of residuals, the motion vectors, the percentage of skipped macroblocks, the percentage of quantized zeroes, etc.), in other words, two different sequences in terms of the objects' motion speed in the scene, would be encoded with the same λ for a given QP. It has the drawback of considering the macroblocks as being identical from the standpoint of the statistical content. The standard encoding method uses λ_{HR} as described in equation 1.6 is not regarded as optimal because it was determined while looking for the minimal distortion, which only occurs at higher bit rates. It is only at higher rates that λ (which converges to λ_{HR}) depends asymptotically on QP only. However, it might be more efficient to encode with a distortion just a bit higher and have the benefit of a much lower bitrate, if possible. We can look for values of λ that are more appropriate for the encoding of the lower bit rate, typically represented by slow-paced sequences and the only solution is to relate λ, besides QP, to the statistical characteristics of the input content that may dictate the value of λ too. Encoding the slow-paced sequences with

adaptive values of λ other than λ_{HR} may result in lower bitrate at the same video quality or even better.

Since the video sequences are so different in any terms and features, the current expression of $\lambda_{MODE} / \lambda_{MOTION}$ is a wasteful approach in terms of bit allocation. An adaptive λ with the nature of the sequences, able to be used at lower bitrate too, would be more appropriate, necessary, and sustainable with the ongoing technological progress, since Lagrange multipliers will feed on these properties as they become available during the sequence encoding. The statistical features and the probabilistic nature of the video sequence are the key of the new optimal compression algorithms.

1.2.3 Lagrange multiplier for high rate encoding

An old-fashioned, empirical Lagrangian multiplier valid at high rates only is used in encoding. It is employed at low rates too, even though the low rate region is unstable and less predictable than the high rate domain. Consequently, there are no models for this region, whatever the sequence type (fast, medium or slow). The expressions of rate, distortion and Lagrange multiplier for the video compression standard H.263 was outlined in the articles (Wiegand and others, 2003), (Sullivan and Wiegand, 1998),and (Wiegand and Girod, 2001). The quadratic dependency in H.263 was replaced by H.264's exponential behavior but the dependency is still static with any sequence, as outlined in Table 1.2.

Table 1.2 Lagrange multiplier with high rate assumption

	Distortion	Rate	λ_{MODE}	λ_{MOTION}
H.263	$D = \frac{Q^2}{3}$	$R(D) = a \ln(\frac{\sigma^2}{D})$	$0.85Q^2$	$\lambda_{MOTION \mid SSE} = \lambda_{MODE}$ $\lambda_{MOTION \mid SAD} = \sqrt{\lambda_{MODE}}$
H.264/ AVC			$(0.85).2^{\frac{QP-12}{3}}$	

1.3 General block diagram of video compression

A video encoder is a system that receives the video source as input and outputs an approximation of it in order to deliver the minimum amount of bits that still maintains a high quality of the image. Its components, featured in the previous subchapters, may be grouped according to (Richardson, 2010) into several models, of which, the most important are the prediction model (spatial and temporal), the image model (predictive coding, transform coding and quantization), and the entropy encoder.

In a more detailed picture, the encoder in Figure 1.8 contains some of the elements of the RDO mechanism as well. The video source is represented by the current frame, from which the current MB is selected for encoding. The prediction model, which exploits both temporal and spatial redundancy, finds the best estimation (multiple prediction blocks when MB is partitioned as in Figure 1.3) from the previously encoded frames, and partially removes the redundancy between the original macroblock (MB) and the estimated one in spatial and temporal domains. This is where the RDO mechanism, by means of λ_{MOTION} of the equation (1.12), trades the number of bits to encode the MVs for the distortion calculated between the current and estimated MBs. Once the best prediction and MVs are found, the residual signal as a difference between the original MB and its prediction, motion compensated, is fed to the image model. The residual is transformed, quantized, inverse quantized and inverse transformed to produce the decoded residual which is added to the prediction to form the reconstructed MB. The optimal mode is concurrently decided via λ_{MODE} of the equation (1.13) that helps negotiate between the number of bits to encode the transformed and quantized coefficients, and the distortion D(mode), calculated between the reconstructed an the original MBs. The bits required to separately encode the MVs difference and the transformed/quantized coefficients of the best mode, are included into the final bitstream.

In the current reference implementation of H.264, the block that generates the Lagrangian multipliers for the ME and MD stages, uses empirical values and is totally independent of any of the processing blocks in the diagram. In this research, we improve the design by

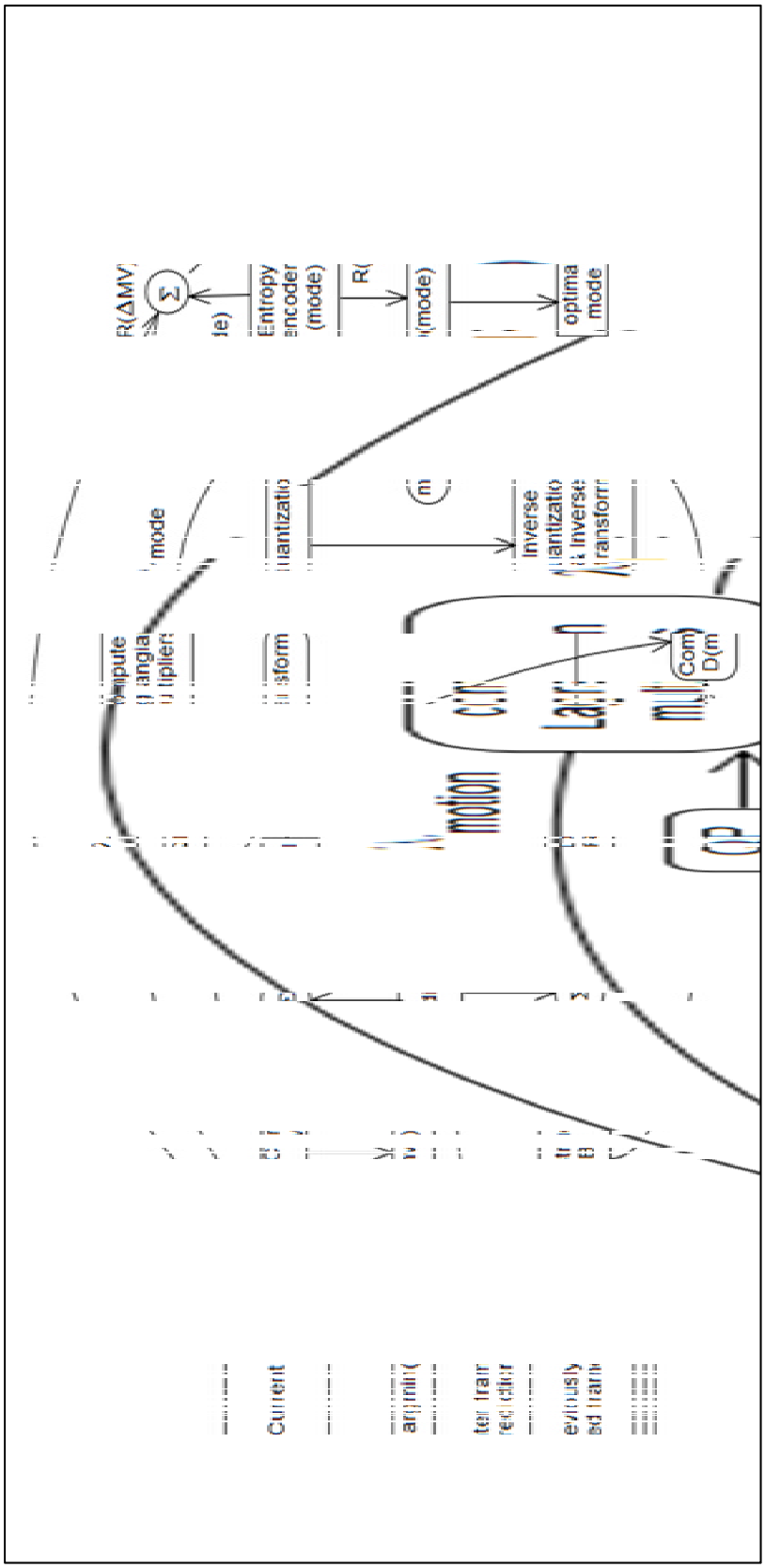


Figure 1.8 H.264 encoder block diagram.
Adapted from (Richardson, 2010)

connecting it to the transform block of the mode decision loop and propose a new algorithm to dynamically adapt the Lagrange multipliers with the image content, thereby improving the performance in terms of rate-distortion.

CHAPTER 2

LITERATURE OVERVIEW

2.1 Lagrange multiplier selection

In H.264/AVC standard the Lagrange multiplier selection occurs at both motion estimation and mode decision levels, this is why it has 2 components: λ_{MOTION} and λ_{MODE}. Yet, only when the SSE metric is utilized in both processes these values are deemed equal, which is in line with the way the distortion (PSNR) is calculated. The value of λ_{MOTION} displays a weak dependency on the search precision (FPel, HPel, QPel) and method (full search, UMHexagon, EPZS with its refinement patterns) utilized in motion estimation. The best prediction of a macroblock, once established, is used throughout the mode decision process to establish the best trade-off between rate and distortion achievable for a certain encoding mode. Even with a suboptimal prediction (generated by a sub-optimal search method) the mode decision is the one that finally decides what is the best mode for a macroblock to be encoded with, so the mode decision outweighs in importance the motion estimation stage. It becomes stringent that the value of Lagrange multiplier is the right one, especially for the mode decision stage; and this is an area where the research has been focusing in the latest decade.

Currently, the reference software (Sühring, 2013) allows the encoding of the current MB by using up to 16 frames as reference. A frame is encoded using one of the patterns I, P, B. To encode a P frame, each MB, based on the RDO mode decision, (High, High Fast, and Low) can be encoded as 16x16 8x16 16x8, 8x8 and each block from an 8x8 partition can be encoded as an 8x4, 4x8 or 4x4 block. To encode a 4x4 block pertaining to I-type frame there are 9 possibilities based on the samples supplied by the neighborhood.

Besides, in the process of the motion estimation, there are multiple ways of finding the best match for a MB given the reference frame(s).

When the discrete J cost function is calculated for a MB all these possibilities come into play, and the optimum value is the one for which the minimum of J is achieved among the entire set of possible configurations.

The H.264/AVC reference software implements a one-step encoding algorithm at the macroblock level, where the value of Lagrange multiplier is calculated by taking into account the rate and distortion dependency on the quantization parameter QP only.

Even so, the experiments on multiple sequences, with slow and fast movement, along with the usage of different combinations of profiles and types (baseline, main, extended) have proven that the rate and distortion models (or similarity as a measure of visual conformance) also depend on many other factors among which the most important are the side information (macroblock header bits, MV bits, frame and macroblock type, entropy model), the source information, selection of the encoding modes, and especially the information contained in the transformed coefficients, as pointed out in (Li and others, 2009) and (Li and others, 2007).

Indeed, more accurate rate and distortion models would have to consider these parameters as well. The calculation of the Lagrange multiplier would then have to take into account the partial derivatives with respect to all these parameters, for both continuous and discrete cases. In this way, such models would get even closer to real data. Yet, an obvious downside would be the increasing computational effort that would occur in this case. This is why, in practice, the encoders follow the Wiegand-Girod RD model that depends on QP only (Wiegand and Girod, 2001).

A quadratic or exponential dependency with QP is expressed in the case of H.263 or H.264/AVC respectively. The rate-distortion model of Wiegand-Girod is based on the high rate assumption, which allows expressing the Lagrangian multiplier λ_{HR} in terms of QP at high rates. Any Lagrange multiplier expression, determined for any other model, should asymptotically converge to this value, as a measure of the new approach correctness. As such, the H.263 and even H.264/AVC adopted the λ_{HR} approach, regardless of the motion

degree in the scene, which experimentally proved satisfactory in combination with the traditional distortion measures SAD, SATD, and SSE. Still, the high rate assumption might not hold true in the case of video conferences or slow-paced sequences (ex: container.yuv, bridge.yuv), where unnecessary bits might be sent into the resulting bitstream.

Since λ_{HR} is only related to QP, as long as the other dependencies mentioned beforehand are neglected, λ is not optimally calculated and the calculated values of the cost function J are higher, so the performance is weak.

The main drawback of the new models developed so far, resides in the fact that within the calculation of the Lagrangian multiplier, the derivatives of the rate and distortion are computed with respect to Q step (represented by the QP parameter) only.

Several trends related to bits allocation have been emerging in the recent years. Some of them try to fit the best probability distribution function with the residual signal distribution. Others replace the distortion with its complementary, which is based on a perceptual visual quality metric, to explore for the optimal expression of λ, or they combine the compression processes at the frame and macroblock levels successively optimizing for one parameter at a time.

Their performances, benefits, and drawbacks are featured in the following sections.

2.1.1 Laplace distribution-based approach for inter-frame coding

The articles (Li and others, 2009) and (Li and others, 2007) constitute the first step taken to model the rate and distortion functions from other perspective, beyond the traditional sole dependence of QP. The novelty of the proposed method consists in a new Lagrange multiplier determined at the frame level, equally applicable to all frame's macroblocks, and adaptive with the frame's statistics. The algorithm efficiency was proved especially in the case of the quasi-stationary sequences, though the authors needed to handle several special

cases where the theoretical assumptions and the proposed model did not fit with the real-world use.

There are several great contributions and interesting standpoints in this article, as follows:
1) The derived RD models are based on the zero-mean Laplace distribution of the coefficients resulting from the transformed residuals. The Laplace parameter Λ and the standard deviation σ of the transformed residuals are related as follows:

$$\Lambda = \frac{\sqrt{2}}{\sigma} \tag{2.1}$$

The zero-mean Laplace distribution is defined as:

$$f_{Lap}(x) = \frac{\Lambda}{2} e^{(-\Lambda|x|)} \tag{2.2}$$

The Laplace distribution was chosen among other distributions (Cauchy, Gauss) due to its single parameter Λ to be determined. It also has a good accuracy and a medium complexity of the calculation. The hypothesis of Laplace distribution of the transformed residuals overrides the high rate assumption, for it can be applied for low rate output too. Likewise, the standard deviation σ of the transformed residuals is strongly related to the source of video signal, being considered an inherent statistical property of the input sequence. Consequently, the Laplace distribution establishes itself as a unanimously agreed-upon choice for the representation of the input signal distribution.

2) The entropy H of the quantized transformed residuals is calculated based on the uniform reconstruction scalar quantizer, though, being dependent on the encoding method (CAVLC or CABAC), its expression roughly represents the rate model. The authors of (Li and others, 2009) refined the expression of the entropy, obtained on the following considerations:
- The probabilities of the transformed residuals, summed as the entropy, are calculated by integrating the Laplace pdf of the uniform quantized residuals, within the quantization intervals, corrected with the rounding offset γ. The offset constant values were separately determined for intra/inter encoding, though it could itself constitute another parameter for performance improvement.

- The classical form of the entropy ($H = -\sum P.\log_2 P$) was corrected in order to handle the case of skipped macroblocks and get as close as possible to the real rate.

- Since the entropy cannot further handle the final stages of the compression (run length and tree/arithmetic encoding), the authors were compelled to apply correction factors to the probability P_0 of quantized-zeroes (computed on the dead zone) and probability P_n of quantized non-zeroes, respectively. The resulting rate model excludes the bitrate of the skipped macroblocks.

- The ratio $r = P_s / P_0$, where P_s is the probability of the skipped blocks and P_0 represents the percentage of the quantized zeroes per frame, is always sub unit. Therefore, this derived parameter is considered as an inherent property of the input sequence too.

- A roughly linear dependency relationship, at the practical QP = 28..40 values, was experimentally noticed between $\ln(R / H_{refined})$ and the product (Λ.Q). The linearity constant S was determined under the convergence condition.

3) The closed form of the distortion model is determined by summing the second moment of the Laplace pdf on each quantization interval.

4) The authors proved that for uniform distribution, which can be obtained from Laplace distribution when $\Lambda \to 0$:

$$\lambda_{Lap} \to \lambda_{HR}(=c.Q^2) \tag{2.3}$$

as in the case of H.263. It means that when the Laplace signal extends to all frequencies spectrum ($\sigma \to \infty$), λ_{Lap} becomes a particular case of λ_{HR}.

5) This is an adaptive algorithm whose parameters values (Λ, r) in the current frame are estimated from the ones collected in the previous (up to five) frames. With the predicted

values, the value of λ_{Lap} can be calculated from the derivatives of rate and distortion models with respect to quantization step.

Finally, the Lagrange multiplier is calculated as

$$\lambda_{Lap} = -\frac{\partial D}{\partial R} = -\frac{\partial D/\partial Q}{\partial R/\partial Q} \tag{2.4}$$

The approach of Laplace-based rate distortion models has some limits, though.

Firstly, the derived rate model does not consider the side information (MV, MB type). Fortunately, the effect of side information was taken into consideration in the escape methods that accompany the algorithm for the cases that cannot be captured by the RD models.

Secondly, the algorithm considers that a single Laplace distribution applies to the whole frame while different regions of the frame may have different Laplace parameters. Estimating the Laplace parameter by taking advantage of each macroblock configuration might be more effective.

The core algorithm and rate distortion models are derived based on several assumptions:
1) Laplace distribution of transformed residuals was the sole distribution considered
2) No side information
3) No sharp scene changes (only smooth scenes are allowed)

In order to manage the encoding of sequences with a large range of characteristics the algorithm was supplemented with escape methods (corrections) to cope with the situations when the assumptions above do not hold valid in the real world. A correction process is initiated once an assumption is detected as invalid for the current frame. The correction is considered as an adaptive process guided by a self-corrector algorithm.

In order to detect sharp changes in the current frame, the standard deviation σ_0 of the current frame with respect to the previous same-type reconstructed frame is employed, and an adjusting process for λ_{Lap}, based on several constants experimentally determined, takes place. When a non-Laplace distribution is detected in the transform residuals of the input sequence, the index

$$RD_{Lap} = \frac{R_{real}.D_{model}}{R_{model}.D_{real}} \tag{2.5}$$

is employed in order to appraise the proper weighting between rate and distortion calculated using the model equations, and their real coded values in the frame. Then, the algorithm provides formulae to clip the λ_{Lap} value if a continuous non-Laplace (over the latest few frames) occurred or just the current frame did not meet the Laplace distribution requirements. In order to manage the side information (header, MV, macroblock type) in the derived models, correction indexes (gaps between the model and the real world) for both rate and distortion are defined.

R_{gap} measures the ratio of the difference between the rates of residuals with zero motion $R(\Lambda^i{}_0, r^i, Q^i)$ and with motion compensation $R(\Lambda^i, r^i, Q^i)$, and the real rate of the side information $R^i{}_{rs}$. The Laplace parameter $\Lambda^i{}_0$ is calculated for residuals with zero motion.

$$R^i{}_{gap} = \frac{R(\Lambda^i{}_0, r^i, Q^i) - R(\Lambda^i, r^i, Q^i)}{R^i{}_{rs}} \tag{2.6}$$

D_{gap} is calculated as the relative gain of the distortion calculated with zero motion $D(\Lambda^i{}_0, Q^i)$ over the one that takes into account the motion compensation $D(\Lambda^i, Q^i)$.

$$D^i{}_{gap} = 10\log_{10}\frac{D(\Lambda^i{}_0, Q^i)}{D(\Lambda^i, Q^i)} \tag{2.7}$$

The empirically determined thresholds are used to assess the effect of the side information, either for fast/complex scenes or quasi-stationary sequences. Each evaluation of the three assumptions is followed by a refresh process (clipping) of the λ_{Lap} back to the traditional λ_{HR} as the last solution.

In conclusion, the Laplace based RDO is an adaptive method to compute the Lagrangian multiplier at the frame level, based on the information in the input sequence, that works fine under certain conditions: Laplace distribution of the transformed residuals, no scene change and no side information in the bitstream.

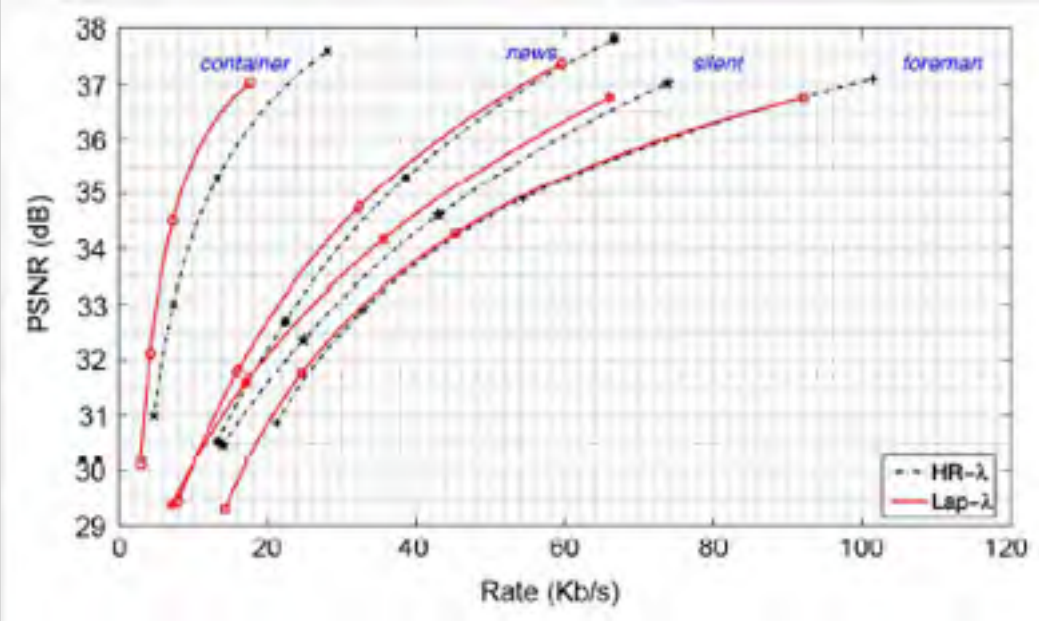


Figure 2.1 RD curves, from (Li and others, 2009)
with permission granted by IEEE

However, when these assumptions are not met, the algorithm has the necessary escape techniques (λ_{Lap} refresh technique) to switch back to the traditional Wiegand-Girod lambda multiplier λ_{HR} .

On average, the algorithm achieves gains of 0.34dB in PSNR or 8.23% rate reduction as illustrated in Figure 2.1. A single sequence, container.yuv, benefits massively from it, with a gain in PSNR of 1.79dB and 32% rate reduction.

There is the inconvenient fact that the algorithm does not have great results for all sequences; namely, even at the frame level the distribution of residuals is clearly of Laplace type, the algorithm is however applied at macroblock level, whose distribution is not always Laplace. This is why, in order to avoid drift, the algorithm contains correction steps.

2.1.2 Laplace distribution-based approach for intra-frame coding

In (Li, Oertel and Kaup, 2007) the same authors of (Li and others, 2009) and (Li and others, 2007), proposed an adaptive Lagrange multiplier selection, this time for intra frame coding. The same R and D models deducted in (Li and others, 2009) were employed.

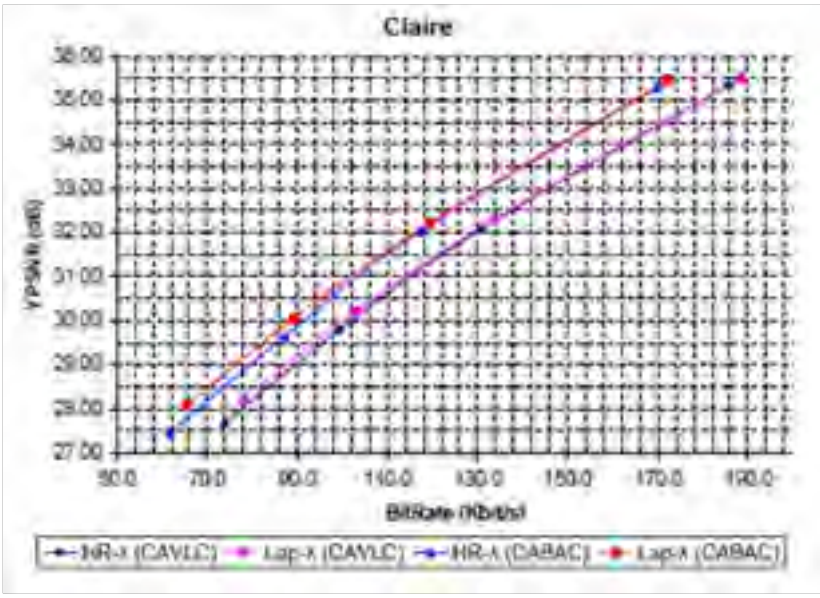


Figure 2.2 RD curves for Claire_qcif.yuv, from (Li, Oertel and Kaup, 2007) with permission granted by IEEE

A Laplace-based Lagrangian multiplier λ_{Lap} was derived according to the variance σ of the transformed residual coefficients. However, for the first frame, entirely intra-frame coded, λ_{HR} was used since no variance information was available before the first frame. The difference from the algorithm presented in (Li and others, 2009) reside in the fact that the statistics regarding RD_{Lap}, R_{gap}, and D_{gap} that would otherwise detect a non-Laplace distribution of the transformed residuals and evaluate the importance of the side information are not used. Nor the escape methods in (Li and others, 2009). Nevertheless, as in (Li and others, 2009), the Laplace distribution parameter Λ gets its current value as the arithmetic mean of the values recorded in the previous frames. The methods achieved gains of up to 0.3dB in PSNR, as illustrated in Figure 2.2.

2.1.3 SSIM-based approach

Since the distortion calculated using the traditional measures (SSE, SAD, SATD) did not prove to be related to the human perception (Wang and Bovik, 2006), the research focused on other measures, capable of being better indicators of the perceived image quality. In this way one can also avoid the use of the PSNR indicator (also related to the conventional metric SSE) to compare the quality of the rendered sequences, since its values domain is unlimited, and as result, it does not have a stated value or interval where the sequence quality might be considered optimal.

SSIM is the indicator that successfully replaces the widely criticized distortion term (SSE) in the functional J, defined as the unconstrained RD cost function to be minimized:

$$J = (1 - SSIM) + \lambda . R \tag{2.8}$$

Although its computation is more elaborate, SSIM takes into consideration the perceptual properties of the image - contrast, luminance, and structure – being considered an objective quality measure.

It adds visual perception levels to the simple mathematical calculation of the error, by considering visual properties (contrast, saturation, structure) at which the error can be analyzed.

SSIM is calculated like its peer PSNR on all components Y, U, and V, of the image decomposition. The overall SSIM is computed by the authors of (Wang and others, 2012) as:

$$SSIM = w_Y * SSIMY + w_U * SSIMU + w_V * SSIMV \tag{2.9}$$

with the weights $w_Y = 0.8, w_U = w_V = 0.1$.

SSIM is totally adaptive to the reference signal according to (Wang and others, 2004). The value of the SSIM index at image level is calculated by averaging local SSIM values obtained using a sliding window at both the reference and distorted images.

Apart from its incontestable qualities in the detection of the image quality, the SSIM index has several drawbacks. Firstly, it is a spatial static indicator since it is calculated at the current image level and has yet to provide a hint with regard to the former or future image quality, as it is the case with the inter frames (P or B). It faces the same issue with the denominator as PSNR does when the error converges to 0; fortunately it uses several empirical constants, to avoid the singularities.

Though the previous RD models and Lagrangian multipliers were derived without deeming the perceptual qualities of the input sequence, the current research is focused on perceptual RD models, based on perceptual measures.

The articles (Wang and others, 2012) and (Wang and others, 2011) would represent the state-of-the art for the next generation of encoders by the novel approach they propose.

The main contributions of (Wang and others, 2012) and (Wang and others, 2011) are:
1) The best mode selection to encode the current MB, which results as a trade-off between the distortion value and the number of necessary bits, represents the goal to be achieved by the encoder. As a result, in the first step, the calculation of λ_{SSIM} is performed at the macroblock level, with surrounding pixels participating as an extension of the current MB, to avoid discontinuities at the original macroblock borders and to provide more samples for calculation, hence increasing the method stability. This situation happens when the marginal pixels of the original MB participate in the calculation of the parameters (μ_x , μ_y , σ_x , σ_y , σ_{xy}) for the current position of the 4x4 sliding windows, in both original and reconstructed frame. Typically, three pixels wide extra side bands are added for luma and chroma components. Thus, the MB's distortion term of the functional J is calculated by means of SSIM.

2) Prior to calculate λ_{SSIM} , SSIM and rate models are derived at frame level. As in the case of (Li and others, 2009), the same approach, based on the Laplace distribution of the

transformed residuals is chosen to calculate the Lagrangian multiplier at the frame level, as a first approximation of λ_{SSIM} which is further refined at the macroblock level.

3) Since the distorted frame is not available to calculate the full-reference SSIM model as per definition, a reduced-reference model is derived. Typically, instead of calculating

$$SSIM_{FR} = \left\{1 - \frac{\left[X(0) - Y(0)\right]^2}{X(0)^2 + Y(0)^2 + N.C_1}\right\}\left\{1 - \frac{\sum_1^{N-1}\left[X(k) - Y(k)\right]^2}{\sum_1^{N-1}\left[X(k)^2 + Y(k)^2\right] + N.C_2}\right\} \tag{2.10}$$

using the definition of SSIM evaluated between the original **x** and the distorted **y** 4x4 blocks, the DCT coefficients are computed at the block level, with the same frequency coefficients grouped into 16 subbands. The mean μ_i and standard deviation σ_i are estimated for each subband *i* of the current macroblock.

The total reduced-reference SSIM measure M_{RR} is computed as a product of reduced-reference SSIM indexes of each AC frequency *i* through the reduced-reference SSIM corresponding to the DC coefficients.

$$M_{RR} = \left(1 - \frac{D_0}{(2\sigma_0^2 + C_1)}\right)\left(1 - \frac{1}{N-1}\sum_1^{15}\frac{D_i}{(2\sigma_i^2 + C_2)}\right) \tag{2.11}$$

In the final expression the coefficients D_i are calculated as the second moment of the Laplace pdf over all quantization intervals, as in the case of Laplace distribution-based Lagrangian multiplier described in (Li and others, 2009). D_i is a function depending on Λ and Q. M_{RR}'s design is based on the features (σ_i) extracted from the original frames in the DCT domain and from residuals.

A nearly perfect linear dependence between $SSIM_{FR}$ and M_{RR} allows to predict SSIM based on M_{RR}. However, the abscissa's intercept (M_{RR}) is sequence dependent and is dynamically estimated from the previous frame DCT transform i-th subband, while the convergence point (SSIM, M_{RR}) is always equal to (1,1).

4) The rate model is based on the entropy model H, described in (Li and others, 2009) and does not take into consideration the bitrate of the skipped blocks, since the skipped blocks are not coded in the final stream. As in (Li and others, 2009), the linear relationship between $\ln(R / H_{refined})$ and the product (Λ .Q) is fully exploited, by taking into consideration the side information. The side information becomes important in low bitrate; the header bitrate is proportional with the source bitrate. As in (Li and others, 2009), the rate model is proportional with the entropy model and exponentially increases with the product (Λ .QP), where Λ describes the intrinsic properties of the source.

Finally, λ_{SSIM} is calculated by minimizing *J*:

$$\lambda_{SSIM} = \frac{dSSIM}{dR} = \frac{\partial SSIM / \partial Q}{\partial R / \partial Q} \tag{2.12}$$

The value of λ_{SSIM}, determined at frame level, is adapted with the inherent properties of the source (σ_i).

5) The second novelty of the paper consists in the refreshment of λ_{SSIM} at MB level. Any MB (except the skipped ones) is associated the motion information, which is strongly related to the perception quality, assessed every time by HVS.

The MV are used by motion compensation (MC) to reduce the amount of the bitrate, and build the predicted image, used to construct the residuals, but, according to (Wang and Li, 2007), the *perceptual information content* in MBs is not the same in terms of *motion information content* and *perceptual uncertainty*. The study (Wang and Li, 2007) concluded that the Lagrangian multiplier varies directly proportional with total distortion D and inverse proportional with total bitrate. This remark was used to adjust λ at the MB level. As a result more bits are assigned to MBs whose *perceptual information content* (in terms of MVs and DCT coefficients) is more important but the *perceptual uncertainty* is less significant. The weighting factor η that controls the adjustment of λ is defined as:

$$\lambda_{MB} = \eta . \lambda \tag{2.13}$$

where λ is the value calculated at frame level and:

$$\eta = \sqrt[4]{\omega_{avg}/\omega} \tag{2.14}$$

The spatiotemporal importance weight function was defined as:

$$\omega = I - U = (\varphi \log v_r + \nu) - (\log v_g - \tau \log c + \delta) \tag{2.15}$$

where:

I = motion information content = self-information of the relative motion,

U = perceptual information function,

v_r = relative motion vector,

v_g = global background motion vector.

The constants φ, ν, τ, and δ are the result of the psychophysical research that employs the power-law and log-normal distribution to determine the distribution of relative motion and perceptual uncertainty respectively.

The variable c represents the contrast measure that depends exponentially on MB's (μ, σ). With the assumption that the inherent properties of the input sequence can be considered constant for a short period of time, the authors observed that the parameters Λ and ω_{avg} vary slowly enough to be considered constant. Consequently, they are predicted from the previous frames. For the first frames, when the adaptive Lagrangian multiplier cannot be calculated, the authors proposed a new expression for λ_{HR}, which is determined on the high rate assumption and by using the expectation of the SSIM and uniform probability distribution for MSE:

$$\lambda_{HR} = aQ^2 - bQ^4 \tag{2.16}$$

The rate model proposed in (Wang and others, 2012) and (Wang and others, 2011) is valid for high bitrate; at low bitrate the skipped blocks number increases and the source bitrate tends to zero. The method achieved a gain of (0.002,..,0.01) in SSIM or a 5,…,21% rate reduction, as illustrated in Figure 2.3.

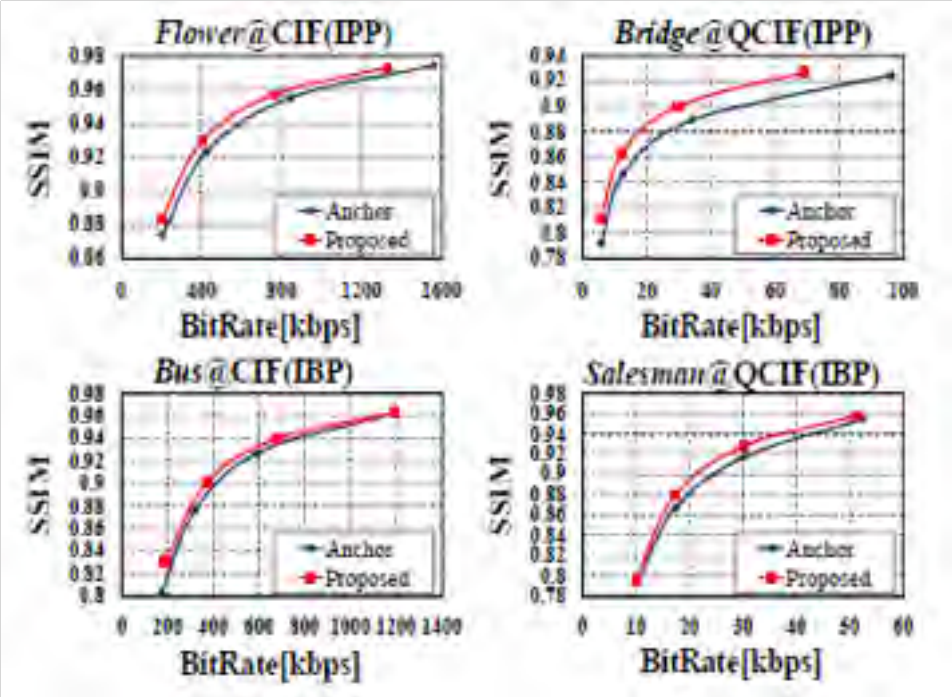


Figure 2.3 RD curves for the SSIM approach, from (Wang and others, 2012) with permission granted by IEEE

As a conclusion, the new methods, based on the various probability distribution functions, in particular the Laplace distribution, seem to be the key approach for the next generation of encoders. The main benefits are the rate reduction corroborated with performance boost and image quality closer to the HVS expectation level.

2.2 Types of distribution of the residual transformed coefficients

The article (Xie and Chia, 2008) is an exhaustive and detailed analysis of the DCT residues from the standpoint of statistical properties. After DCT, the input can be regarded as composite video signal of mostly uncorrelated frequencies. The distribution of DCT residues is dependent of the DCT coefficients and the quantizer characteristics (Q, γ). Beside the DC component, some frequencies contribute, after their quantization, to the final bitrate. According to the authors, Xie, J., and L.T. Chia, the DC coefficients are best modeled by a Gaussian pdf, whereas the AC components are best modeled as a Laplace source.

Both DC and AC coefficients are passed through a uniform distribution quantizer without threshold. Further, new models for rate and distortion are derived analytically. A rate control algorithm based on the derived expressions of rate and distortion is also discussed. The authors claim that their models can predict the possible distribution prior to actual encoding and without using any empirical knowledge. However there is still a problem, namely, only the entropy (but not the final rate) can be predicted by any of these models.

On the same note, E. Lam and J. Goodman, the authors of (Lam and Goodman, 2000), gave an explanation as to why the most suitable distribution for AC DCT coefficients is Laplace and how the width of the AC DCT distributions shrinks at high frequencies while preserving the shape. The energy is smaller in higher frequency subbands and spatial correlations contribute to the decreasing of the distribution width.

There are also other attempts to model the empirical data, made with other distribution types. The authors of (Altunbasak and Kamaci, 2004) proposed the use of zero-mean Cauchy pdf

$$pdf_{Cauchy}(x) = \frac{\mu}{\pi(\mu^2 + x^2)} \tag{2.17}$$

claiming that it offers better fit with the AC DCT residuals than Laplace distribution. Unlike Laplace whose parameter Λ can be easily deduced in relation to the variance σ, Cauchy distribution needs to have the parameter μ dynamically determined for each set of samples, but has the advantage of having simpler approximated expressions for rate and distortion:

$$R(Q) \approx aQ^{-\alpha}, D(Q) \approx bQ^{\beta} \tag{2.18}$$

where $a, b, \alpha, \beta > 0$. Nevertheless those expressions were determined for a rounding offset $\gamma = 1/2$, no matter the frame type.

As a matter of fact, the curves R(Q) and D(Q), which were depicted for a single frame of each encoding type (1I+1P) and a range of quantization steps show Cauchy a perfect fit for intra-compression while for interframe the low rates were advantaged by its use. However, the results were not confirmed for a sufficient greater number of frames to draw a positive conclusion.

Guided by the knowledge acquired from the articles aforementioned, the first step of the research involved the search of the best distribution to associate with the macroblocks DCT residuals. The best candidates to take into consideration were the well-known Laplace and Gauss distributions.

The choice is principally justified by the pdf shape, as noticed from the macroblock pdf graphics in Figures 2.6 and 2.7.

2.2.1 Gaussian model

The reason why Gauss distribution is taken into account is due to its presence at the macroblock level in all motion type sequences, when the criterion (2.31) is applied to each macroblock. The Gaussian model fits macroblocks with symmetrical, mesokurtic distributions of transformed residuals, showing concave shoulders and short, usually slender tails. It appears more frequently at lower QP where the percentage of non-zero-mean transformed coefficients is higher, even for slow-paced sequences, as shown in Figure 4.1.

2.2.1.1 Distribution equation

The complete form of the Gauss distribution is:

$$pdf_{Gauss}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \tag{2.19}$$

where μ and σ are the mean value and standard deviation respectively.

2.2.2 Laplace model

The Laplace distribution can be represented by two side-by-side exponential probability distributions functions. Like the normal distribution function, its shape is symmetric, but with thicker tails and a sharp singularity located on the symmetry axis. In video encoding the

distribution's sharp peak is due to the high percentage of the zero value residues that occur mainly at the encoding of the slow sequences.

The Laplace distribution fits the transformed residuals commonly found in quasi-stationary sequences (container, bridge) and portrait-like sequences with flat background (akyio, miss-america, claire, news) and small degree of motion in the scene. The Laplace distribution is noticed at the transformed residuals quantized with higher QP values, as shown in Figure 4.1.

2.2.2.1 Distribution equation

The complete form of the symmetric Laplace distribution is.

$$pdf_{Lap}(x) = \frac{1}{2\theta} e^{-\frac{|x-\eta|}{\theta}} \tag{2.20}$$

, where η and θ are the location and scale respectively.

The transformed residuals in the majority of the cases have their mean close to zero, which leads to the following form of the Laplace pdf, which is used to compute the distortion, entropy, and Lagrange multiplier:

$$pdf_{Lap}(x) = \frac{\Lambda}{2} e^{-\Lambda|x|} \tag{2.21}$$

where Λ stands for Laplace parameter.

The relation between Λ and σ was determined from the condition of Laplace distribution having the same variation as Gauss distribution, given that their expected values - location (Laplace) and the mean (Gauss) - are the same. Table 2.1 presents a parameter comparison between the Laplace and Gauss distribution.

Even if the location η was stripped out from the general formula of the Laplace distribution, the form (2.20) can be considered quite exact at the frame level, where the number of

Table 2.1 Laplace and Gauss distribution - parameters comparison

Distribution	Variance	Skewness	Kurtosis
Laplace	$2\theta^2$	0	3
Gauss	σ^2	0	0

samples is big enough even for small resolutions i.e. 25,344/1,115,136 samples in the case of QCIF/CIF (Quarter Common Intermediate Format/Common Intermediate Format) resolution respectively. But it might miss accuracy and have the zero-mean assumption denied at the macroblock level and further at its subbands level, where the number of samples cannot surpass 256 and 16 samples respectively.

2.2.3 Generalized Gauss model

The generalized Gauss distribution (GGD), was taken into consideration in the articles (Sun and others, 2013a), (Sun and others, 2013b), and (Zhao and others, 2010), though not for the purpose of the RDO. GGD reduces to Laplace distribution and Gauss distribution when $\alpha = 1$ or $\alpha = 2$ respectively as illustrated in Figure 2.4.

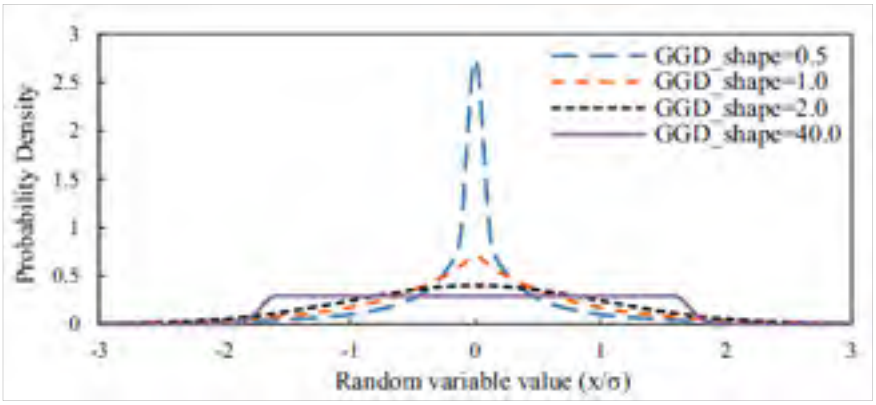


Figure 2.4 Zero-mean generalized Gaussian distribution.
Adapted from (Sun and others, 2013a)

This model was considered for the cases where neither Laplace nor Gauss distributions fit the real world, but the real distribution could fit an intermediate shape from the same family of Laplace or Gauss distributions.

2.2.3.1 Distribution equation

The complete form of the zero-mean generalized Gauss distribution (GGD) as specified in (Sun and others, 2013a) is:

$$pdf_{GG}(x) = \frac{g_1(\alpha)}{\beta} e^{\left(-[g2(\alpha)\frac{|x|}{\beta}]^{\alpha}\right)} \tag{2.22}$$

where:

$$g_1(\alpha) = \frac{\alpha.\Gamma(3/\alpha)^{1/2}}{2.\Gamma(1/\alpha)^{3/2}} \tag{2.23}$$

$$g_2(\alpha) = \left(\frac{\Gamma(3/\alpha)}{\Gamma(1/\alpha)}\right)^{1/2} \tag{2.24}$$

The discrete values of the parameters of GGD, α (shape parameter) and β (standard deviation) can be estimated according to the formulas in (Sun and others, 2013b) and (Zhao and others, 2010):

$$\alpha^* = F\left(\frac{\left(\frac{1}{N}\sum_{k=1}^{N}|X_k|\right)^2}{\frac{1}{N}\sum_{k=1}^{N}(X_k)^2}\right) \tag{2.25}$$

$$\beta^* = \sqrt{\frac{1}{N}\sum_{k=1}^{N}(X_k)^2} \tag{2.26}$$

The function

$$F(x) = \frac{0.2718}{0.7697 - x} - 0.1247 \tag{2.27}$$

designates the fit function for α^*, where X_k are the transformed coefficients of the residuals of the macroblock *k* in the frame.

2.2.4 Proof test

The goodness of fit of Laplace distribution was discussed by *Puig P.* and *Stephens M.* in (Puig and Stephens, 2000) and (Puig and Stephens, 2007). The statistical test considers the criterions of Kolmogorov-Smirnov and Cramer von Mises - based on the cumulative distribution function - in order to compare two empirical distribution functions. A far simpler, parameter-based criterion, to choose Laplace over Gauss distribution, was demonstrated by *Kundu D.* (Kundu, 2005), using the ratio of maximized likelihood (RML).

For a number of n i.i.d. samples $X_1,..,X_n$, the maximum likelihood estimators $(\hat{\mu},\hat{\sigma})$ and $(\hat{\eta},\hat{\theta})$ of the characteristics (μ,σ) and (η,θ) of the Laplace and Gauss distributions are calculated according to the following formulae for the estimated values:

$$\hat{\mu}=\frac{1}{n}\sum_{i=1}^{n}X_i \tag{2.28}$$

$$\hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}(X_i-\hat{\mu})^2 \tag{2.29}$$

$$\hat{\eta}=median\{X_1,..,X_n\} \tag{2.30}$$

$$\hat{\theta}=\frac{1}{n}\sum_{i=1}^{n}|X_i-\hat{\eta}| \tag{2.31}$$

The statistical test *T* is based on RML (ratio of maximum likelihood) of the likelihood functions of Laplace and Gauss distributions. If the value of the statistical test $T>0$ then the most likely pdf is Gauss type otherwise Laplace type with the significance level $\alpha = 0.05$.

This criterion was intensively used in our research to decide for the particular case when the distribution generated by the transform residuals may be restrictively considered either of type Laplace or Gauss (closer of either Laplace or Gauss pdf). In this way, the macroblocks

of a single frame were coded either with Laplace or Gauss-derived Lagrange multiplier. All the same, the generalized Gauss distribution was used as an extra analytical comparator.

$$T = \ln \frac{\prod_{i=1}^{n} \frac{1}{\hat{\sigma}\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{X_i - \hat{\mu}}{\hat{\sigma}}\right)^2}}{\prod_{i=1}^{n} \frac{1}{2\hat{\theta}} e^{\frac{-|X_i - \hat{\eta}|}{\hat{\theta}}}} = \frac{n}{2}\ln 2 - \frac{n}{2}\ln \pi + n \ln \hat{\theta} - n \ln \hat{\sigma} + \frac{n}{2} \tag{2.32}$$

2.2.5 Generic model based on numerical integration

Besides the widely known Gauss and Laplace distributions, the numerical integration was considered as a mean to deal with cases where none of the above would fit the distribution shape of macroblock samples, because of the lack of symmetry, or increased presence of the outliers. While the majority of the transform residuals at the frame level clearly fits Laplace distribution, as shown in the Figure 2.5, most of the real transformed residuals distributions

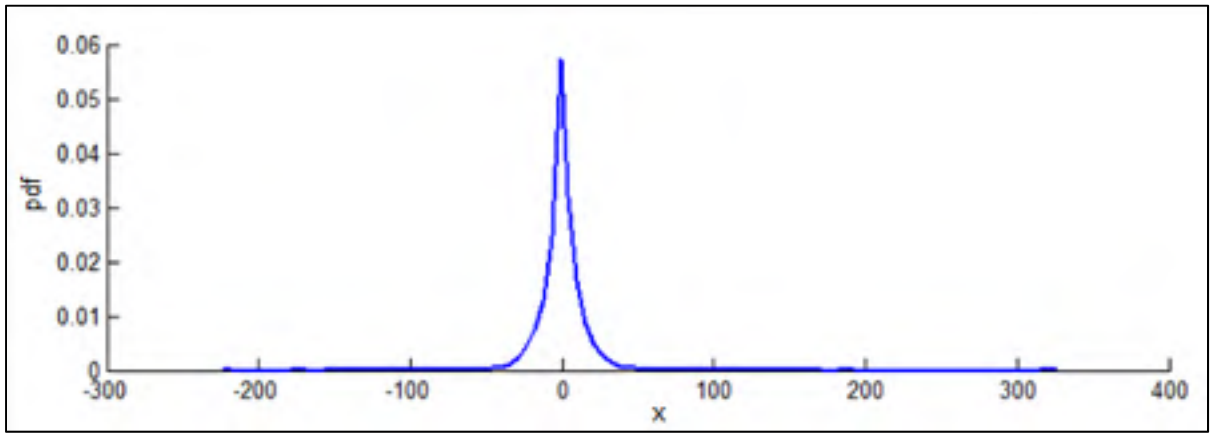


Figure 2.5 The real distribution of transform residuals; seq. Bus (QCIF), frame #6(P)

at the macroblock level do not show a perfect fit with either Laplace or Gauss distribution, though the statistical criterion would indicate the adhesion to the respective probability distribution class.

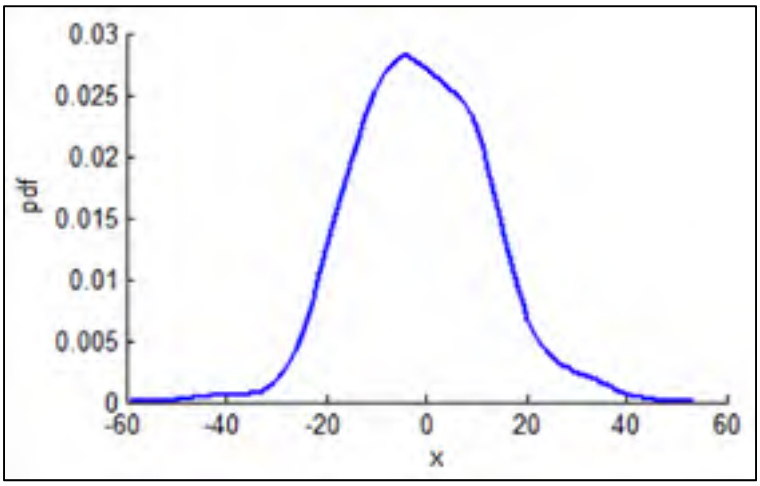


Figure 2.6 The real distribution of transform residuals, frame #6(P), MB (5, 8), sequence bus_qcif

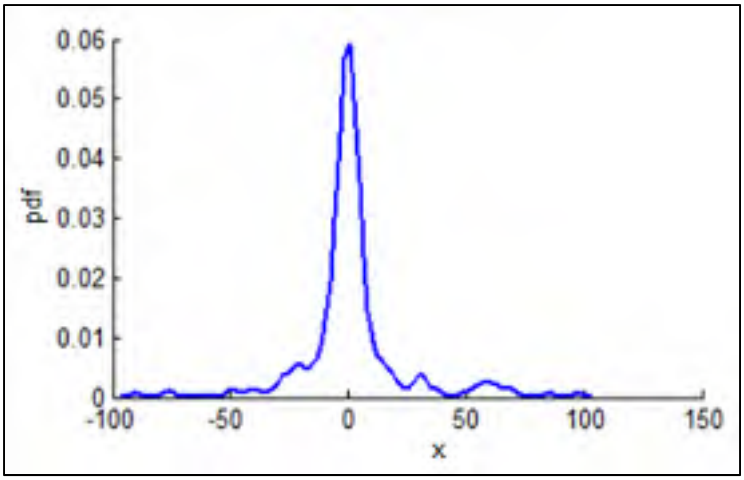


Figure 2.7 The real distribution of transform residuals, frame #6(P), MB (3, 7), sequence bus_qcif

In Figure 2.6 one can see that the macroblock (5,8) would be appropriate for encoding using the Gaussian model formulas, though it lacks the symmetry, while the Figure 2.7 shows the jagged shoulders of the real distribution macroblock (3,7) whose statistical test would entitle it for encoding with the Laplace-based approach.

2.2.5.1 Distribution equation

In practice, the well-known analytical models of Laplace, Gauss and even generalized Gauss do not fit many of the sequence's inputs. In this case, it is better to consider a numerical integration of the distributions generated by the transformed coefficients of the residues. A simple integration method like the trapeze or Simpson approach would deliver an accurate result if the integration step is chosen sufficiently small without affecting too much the computation time and method error. Matlab's ksdensity (kernel smooth density) function generates a sorted collection of discrete pairs $\{x_i, y_i\}$ of the generic probability density function. The collections are utilized in the calculation of distortion, entropy and Lagrange multiplier.

CHAPTER 3

RATE DISTORTION ESTIMATION ASSOCIATED TO LAPLACE, GAUSS, GENERALIZED GAUSS, AND NUMERICAL INTEGRATION COEFFICIENT MODELS

In this chapter we calculate the equations of entropy (which is linked to rate) and distortion for the Laplace, Gauss, and generic coefficient models. Although many of these results are available from the literature, the mathematical derivations are usually lacking. Therefore they are derived here for completeness. Based on these formulas, we make a graphical comparison between the Laplace and Gauss models and their counterparts that form the generalized Gauss distribution.

3.1 General Rate-Distortion equations

When the distribution of the transform residual is known and close to zero-mean, the entropy and distortion can be relatively easily calculated. The general formulas for entropy and distortion contain three terms. The central term is associated to the dead zone $[-Q+\gamma Q, +Q-\gamma Q]$ of the uniform reconstruction scalar quantizer. The other two terms are calculated on symmetric and equally spaced intervals to the left and right side of the dead zone, $[-(n+1)Q+\gamma Q, -nQ+\gamma Q]$ and $[nQ-\gamma Q, (n+1)Q-\gamma Q]$ respectively.

The entropy is calculated according to Shannon's extended formula

$$H = H_n^- + H_0 + H_n^+ = -\sum_{n=1}^{\infty}(P_n^-)\log_2(P_n^-) - (P_0)\log_2(P_0) - \sum_{n=1}^{\infty}(P_n^+).\log_2(P_n^+) \tag{3.1}$$

where:

$$P_0 = \int_{-(Q-\gamma Q)}^{Q-\gamma Q} pdf(x)dx \tag{3.2}$$

$$P_n^+ = \int_{nQ-\gamma Q}^{(n+1)Q-\gamma Q} pdf(x)dx \tag{3.3}$$

$$P_n^- = \int_{-(n+1)Q+\gamma Q}^{-nQ+\gamma Q} pdf(x)dx \tag{3.4}$$

The entropy formula illustrates only the concept of uniform quantization applied to transformed residuals. However, the final steps contained in the entropy encoding block (see Figure 1.1), e.g. the run length and tree (Huffman) /arithmetic encoding, are not addressed by the integration, thus the limitation of the formula that is being used to calculate the rate value as the entropy. Since an analytical formula for the rate is difficult to achieve, at least for now, the values obtained hardly represent the bitrate.

The distortion's general formula:

$$\begin{aligned} D = \sum_{n=1}^{\infty} \int_{-[(n+1)Q-\gamma Q]}^{-[nQ-\gamma Q]} (x+nQ)^2 pdf(x)dx + \\ \int_{-(Q-\gamma Q)}^{Q-\gamma Q} x^2 pdf(x)dx + \\ \sum_{n=1}^{\infty} \int_{nQ-\gamma Q}^{(n+1)Q-\gamma Q} (x-nQ)^2 pdf(x)dx \end{aligned} \tag{3.5}$$

is based on SSE metric and the uniform threshold scalar quantizer with the rounding offset γ at the dequantization (reconstruction).

3.2 Rate-Distortion equations associated to Laplace model

The zero-mean Laplace probability distribution function has the advantage of being an even function, which makes it easy to integrate it using the formulas above. In the case of a zero-mean Laplace-type signal ($\mu \approx 0, \Lambda$) the distortion, entropy and implicitly the Lagrange multiplier depend on Λ, γ, and Q step of the uniform quantizer.

3.2.1 Rate equation associated to Laplace model

The expression of the entropy is:

$$H_{Lap} = \frac{-1}{\ln 2}(1-e^{-(1-\gamma)\Lambda Q})\ln(1-e^{-(1-\gamma)\Lambda Q}) + \\ + \frac{e^{-(1-\gamma)\Lambda Q}}{\ln 2}\left\{\ln 2 - \ln(1-e^{-\Lambda Q}) - \gamma\Lambda Q + \frac{\Lambda Q}{1-e^{-\Lambda Q}}\right\} \tag{3.6}$$

and its demonstration is outlined in ANNEX I. This equation is a simplified version of the one found in (Li and others, 2009) which considers the rate, while we calculate the entropy. When a macroblock was skipped from encoding, a single bit is added to the final bitstream. According to (Li and others, 2009), the relationship between rate and entropy is as follows:

$$R = S.H.e^{-\zeta\Lambda Q} \tag{3.7}$$

where S is a constant derived at the sequence level. (Li and others, 2009) have determined the values of S (1.133 for intra frames and 1.982 for inter frames) from the condition of convergence of $\lambda_{LAPLACE}$, at high rates, towards λ_{HR}. The other constant, ζ, depends on the input video sequence, frame type, and entropy coding method and was experimentally determined by (Li and others, 2009) as 0.35 for CAVLC and 0.30 for CABAC.

3.2.2 Distortion equation associated to Laplace model

For the general case (non-skipped macroblocks) the distortion shows a dependency of Qstep denoted Q and Λ - Laplace parameter.

$$D_{Lap} = \frac{\Lambda Q e^{\gamma\Lambda Q}(2+\Lambda Q - 2\gamma\Lambda Q) + 2 - 2e^{\Lambda Q}}{\Lambda^2(1-e^{\Lambda Q})} \tag{3.8}$$

The distortion of a skipped macroblock $D_{LapSKIP}$, is calculated when the dead zone extends to the whole domain, which corresponds to the case when the quantization does not get applied to the transformed signal.

$$D_{LapSKIP} = \frac{2}{\Lambda^2} \tag{3.9}$$

The demonstration of the distortion formulae is described in ANNEX I.

3.3 Rate-Distortion equations associated to Gaussian model

In the case of a Gauss-type signal (μ , σ) the distortion, entropy, and Lagrange multiplier depend on μ , σ , γ, and quantization step Q of the uniform quantizer.

3.3.1 Rate equation associated to Gaussian model

We calculated the closed form of the entropy H_{Gauss} that we use to express the bitrate. Its demonstration is outlined in ANNEX II.

$$H_{Gauss} = \frac{1}{2}\left\{ erf\left\{\frac{Q(1-\gamma)-\mu}{\sigma\sqrt{2}}\right\} + erf\left(\frac{Q(1-\gamma)+\mu}{\sigma\sqrt{2}}\right)\right\} + \sum_{n=1}^{\infty} P_n^- \log_2 P_n^- + \sum_{n=1}^{\infty} P_n^+ \log_2 P_n^+ \quad (3.10)$$

where

$$P_n^- = \frac{1}{2}\left\{ erf\left(\frac{(n+1)Q-\gamma Q+\mu}{\sigma\sqrt{2}}\right) - erf\left(\frac{nQ-\gamma Q+\mu}{\sigma\sqrt{2}}\right)\right\} \text{ and}$$

$$P_n^+ = \frac{1}{2}\left\{ erf\left(\frac{(n+1)Q-\gamma Q-\mu}{\sigma\sqrt{2}}\right) - erf\left(\frac{nQ-\gamma Q-\mu}{\sigma\sqrt{2}}\right)\right\}$$

As in the case of Laplace distribution, when the macroblock is skipped from the encoding, a single bit is added to the final bitstream.

We have not found any research providing a rate model for the Gaussian case. For simplicity, we assume a similar relationship between rate and entropy as for the Laplace case.

$$R_{Gauss} = S.H_{Gauss}.e^{-\sqrt{2}\zeta Q / \sigma} \quad (3.11)$$

Developing an accurate model would constitute a research project in itself.

3.3.2 Distortion equation associated to Gaussian model

We have not found any research providing a distortion model for the Gaussian case. We calculated the closed form of the distortion D_{Gauss} based on Gauss probability density function.

$$D_{Gauss} = \frac{\sigma^2+\mu^2}{2}\left\{erf\left[\frac{Q(1-\gamma)-\mu}{\sigma\sqrt{2}}\right]+erf\left[\frac{Q(1-\gamma)+\mu}{\sigma\sqrt{2}}\right]\right\}- \tag{3.12}$$

$$-\frac{\sigma}{\sqrt{2\pi}}\left\{[Q(1-\gamma)+\mu]e^{-\left[\frac{Q(1-\gamma)-\mu}{\sigma\sqrt{2}}\right]^2}+[Q(1-\gamma)-\mu]e^{-\left[\frac{Q(1-\gamma)+\mu}{\sigma\sqrt{2}}\right]^2}\right\}+\sum_{n=1}^{\infty}(A_1+A_2+A_3+A_4)$$

where

$$A_1 = \frac{\sigma^2+(nQ-\mu)^2}{2}\left\{erf\left[\frac{(n+1)Q-\gamma Q-\mu}{\sigma\sqrt{2}}\right]-erf\left[\frac{nQ-\gamma Q-\mu}{\sigma\sqrt{2}}\right]\right\}$$

$$A_2 = \frac{\sigma}{\sqrt{2\pi}}\left\{[nQ-Q(1-\gamma)-\mu]e^{-\left[\frac{(n+1)Q-\gamma Q-\mu}{\sigma\sqrt{2}}\right]^2}-[nQ+\gamma Q-\mu]e^{-\left[\frac{nQ-\gamma Q-\mu}{\sigma\sqrt{2}}\right]^2}\right\}$$

$$A_3 = \frac{\sigma^2+(nQ+\mu)^2}{2}\left\{erf\left[\frac{(n+1)Q-\gamma Q+\mu}{\sigma\sqrt{2}}\right]-erf\left[\frac{nQ-\gamma Q+\mu}{\sigma\sqrt{2}}\right]\right\}$$

$$A_4 = \frac{\sigma}{\sqrt{2\pi}}\left\{[nQ-Q(1-\gamma)+\mu]e^{-\left[\frac{(n+1)Q-\gamma Q+\mu}{\sigma\sqrt{2}}\right]^2}-[nQ+\gamma Q+\mu]e^{-\left[\frac{nQ-\gamma Q+\mu}{\sigma\sqrt{2}}\right]^2}\right\}$$

The distortion of a skipped macroblock $D_{GaussSKIP}$, is calculated when the dead zone extends to the whole domain, which corresponds to the case when the signal is not quantized

$$D_{SKIP,\,Gauss} = \mu^2+\sigma^2 \tag{3.13}$$

The demonstration is outlined in ANNEX II.

3.4 Rate-Distortion equations associated to generalized Gaussian model

The complete form of entropy and distortion, as they are calculated by the authors of the articles (Sun and others, 2013a) and (Sun and others, 2013b) depend on the quantization step Q, shape parameter α, standard deviation β, and $z = 1 - \gamma$, the dead zone ratio of the dead zone plus uniform threshold scalar quantization with nearly uniform reconstruction quantization (DZ+UTSQ/NURQ).

3.4.1 Rate equation associated to generalized Gaussian coefficient model

For this model, we are using the closed form of the entropy in (Sun and others, 2013b)

$$H_{GG}(Q) = -\frac{\ln(1-e^{-z.\left[\frac{\sqrt{2}.Q}{\beta}\right]^{\alpha}})}{\ln 2} + e^{-z.\left[\frac{\sqrt{2}.Q}{\beta}\right]^{\alpha}} + \frac{\left[\frac{\sqrt{2}.Q}{\beta}\right]^{\alpha}.e^{-z.\left[\frac{\sqrt{2}.Q}{\beta}\right]^{\alpha}}.[z+e^{-\left[\frac{\sqrt{2}.Q}{\beta}\right]^{\alpha}} - z.e^{-\left[\frac{\sqrt{2}.Q}{\beta}\right]^{\alpha}}]}{[1-e^{-\left[\frac{\sqrt{2}.Q}{\beta}\right]^{\alpha}}].\ln 2} \tag{3.14}$$

A reliable model linking the rate and entropy for the generalized Gaussian model is still lacking. As in the case of Gaussian model, we assume a similar relationship between rate and entropy as for the Laplace case.

3.4.2 Distortion equation associated to generalized Gaussian model

The closed form of the distortion is outlined in (Sun and others, 2013b)

$$D_{GG}(Q) = \beta^2 . \left\{ 1 + \frac{e^{-z.\left[\frac{\sqrt{2}.Q}{\beta}\right]^{\alpha}}}{2.\left[1 - e^{-\left[\frac{\sqrt{2}.Q}{\beta}\right]^{\alpha}}\right]} \left(\left[\frac{\sqrt{2}.Q}{\beta}\right]^{2\alpha} .(1-2z) - 2\left[\frac{\sqrt{2}.Q}{\beta}\right]^{\alpha} \right) \right\} \tag{3.15}$$

3.5 Rate-Distortion equations associated to generic model

The closed forms of the discrete formulas of distortion and entropy at the macroblock level were deduced with respect to their continuous forms under the hypothesis of zero-mean approximation.

The integration step was chosen sufficiently small in order to provide the most precise results possible while avoiding the accumulation of rounding errors.

3.5.1 Rate equation associated to generic model

The discrete form of the entropy follows the same approach as in the continuous case.

$$H_{\text{int}} = -\sum_{n=-\infty}^{\infty} P_n . \log_2 P_n \tag{3.16}$$

with

$$P_n = \sum_i y_i . h_i \tag{3.17}$$

where $h_i = x_{i+1} - x_i$ and $x_i, x_{i+1} \in [nQ - \gamma Q, (n+1)Q - \gamma Q]$

The term P_n is calculated as the sum of products of the pairs $\{x_i, y_i\}$ generated by the Matlab's ksdensity function from the original vector of transform residuals where the

quantization intervals are evenly spaced as an effect of using the uniform threshold scalar quantizer. As in the previous cases, we assume linear relationship between rate and entropy.

3.5.2 Distortion equation associated to generic model

The distortion formula is a double summation over the product of SSE differences and values of the generated pdf pairs $\{x_i, y_i\}$.

$$D_{\text{int}} = \sum_{n=-\infty}^{\infty} \sum_{i=1}^{N} (x_i - n.Q)^2 . y_i . h_i \tag{3.18}$$

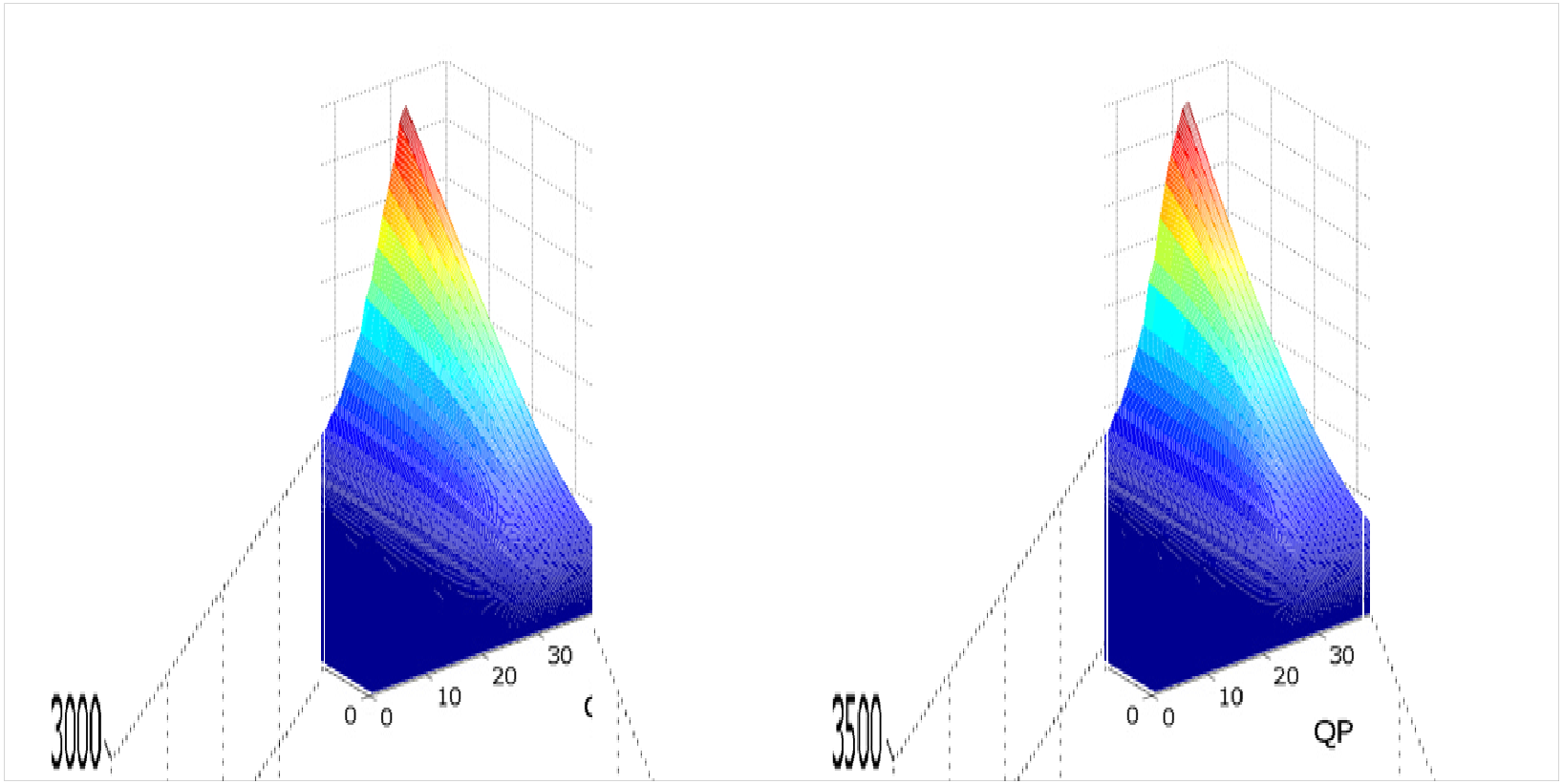
where the integration step $h_i = x_{i+1} - x_i$

In practice, the value of n was restricted to the valid values of abscissae x_i, where the pairs $\{x_i, y_i\}$ were generated by kernel smooth density function.

3.6 Comparison of the distortion models at different QPs

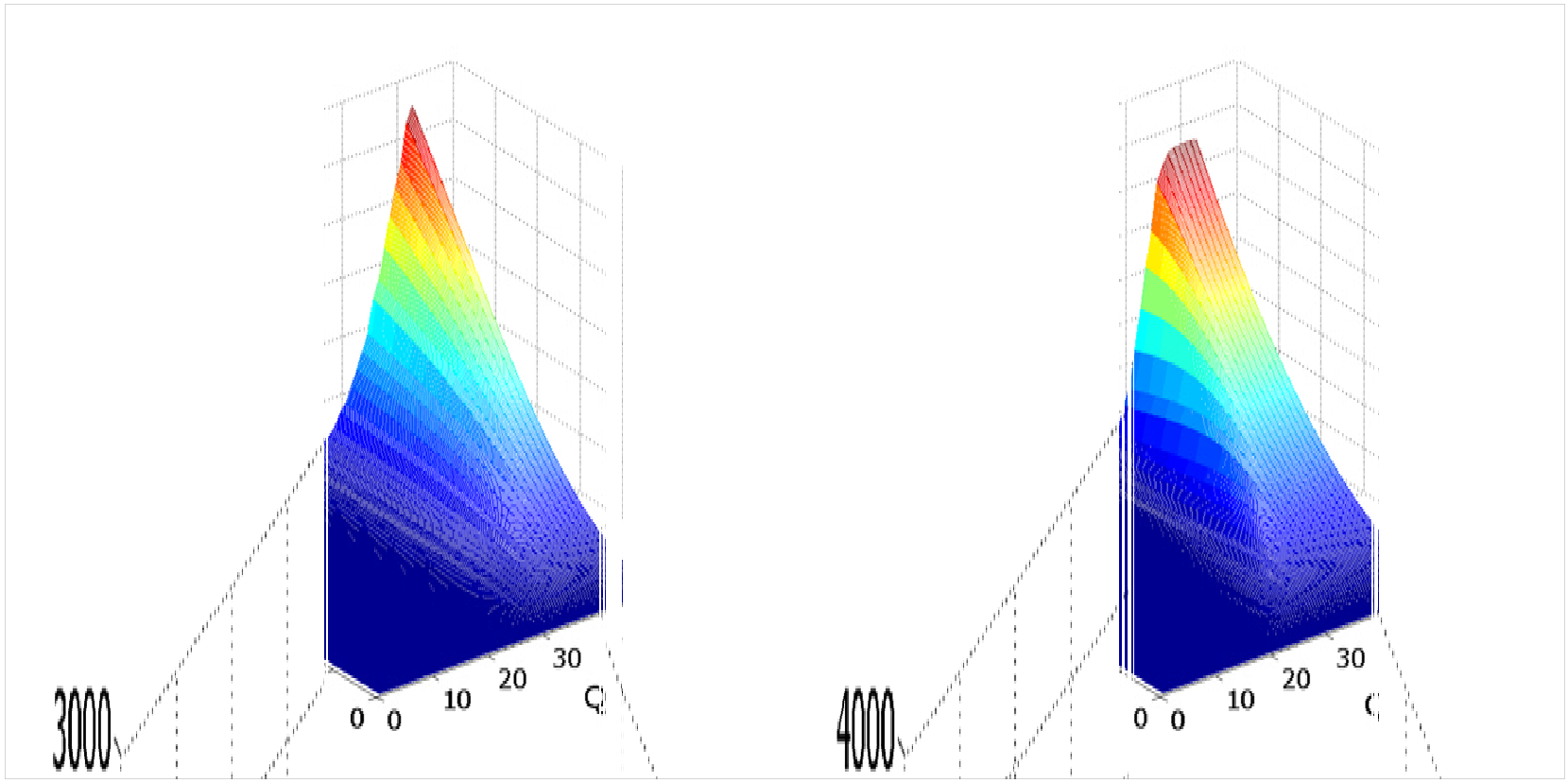
We assessed the range of values the distortion is capable to deliver through analytical formulas. Besides QP, the sequence's intrinsic properties μ and σ at the macroblock level, were considered. The value of μ was set to zero for these graphs, while the interval of σ, (0.06 – 58.1) was determined by collecting these values from a dozen of sequences with slow, medium, and fast scene changes, encoded with the JM's (reference software) standard approach, where the parameter UseCustomLM of the configuration file (ANNEX VI) was set to 1.

The Laplace model can estimate a distortion in the range (0.07 - 2814) while the Gaussian model estimation values are situated in the interval (0.08 – 3312) as illustrated in Figure 3.1. These results were confirmed by their counterparts in the generalized Gauss model described in (Sun and others, 2013a) and (Sun and others, 2013b). The distortion calculated with the generalized Gauss model for $\alpha = 1$ (which emulates Laplace pdf) showed identical range of values as previously determined with the original form of Laplace. All the same, GGD with



a) b)

Figure 3.1 The distortion calculated with: a) Laplace pdf, b) Gauss pdf



a) b)

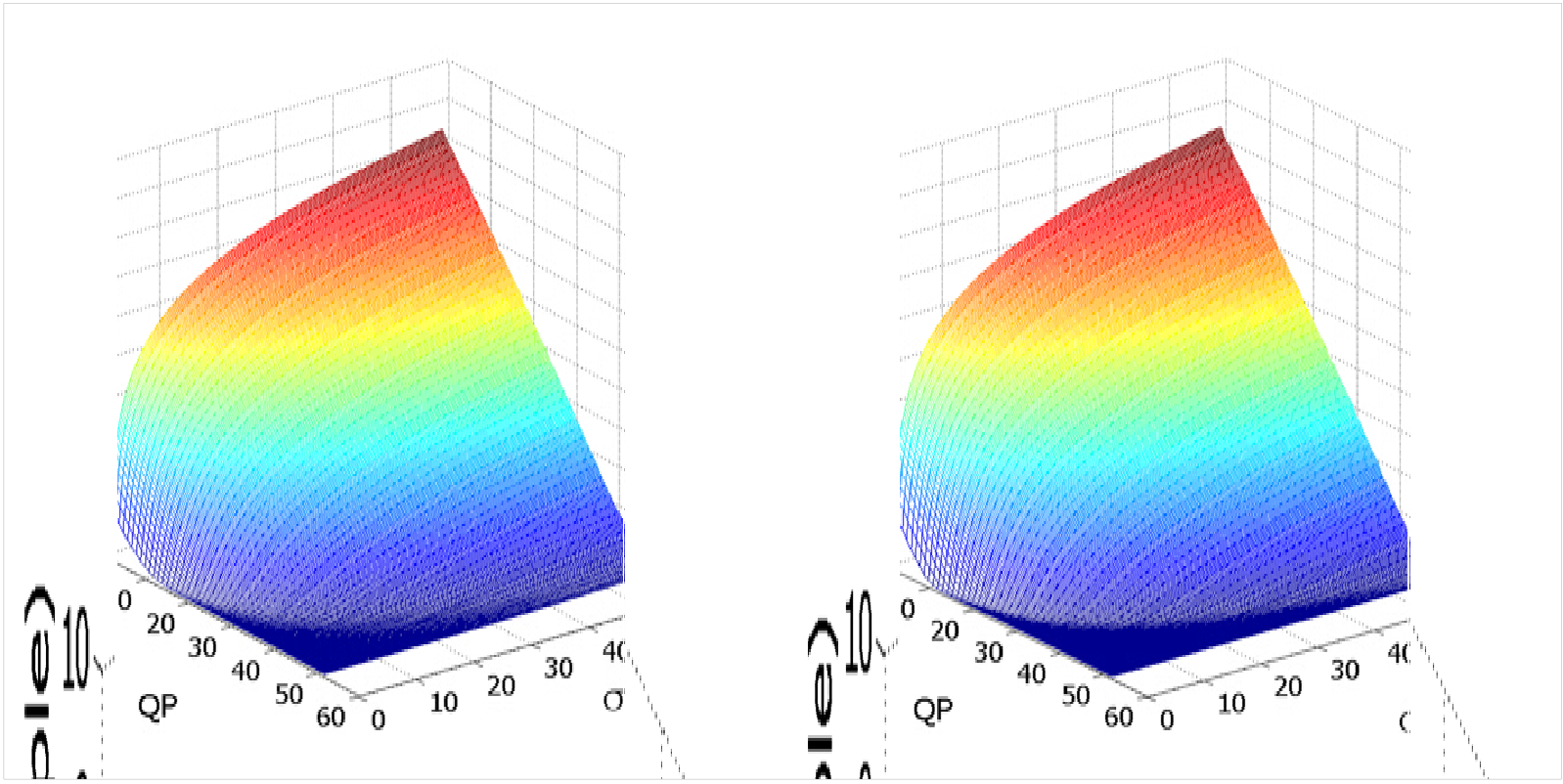
Figure 3.2 The distortion calculated with generalized Gauss: a) $\alpha = 1$ (Laplace), b) $\alpha = 2$ (Gauss)

$\alpha = 2$ produced the interval (0 - 3364), very close to the one generated by the original Gauss distribution, as shown in Figure 3.2.

3.7 Comparison of entropy models at different QPs

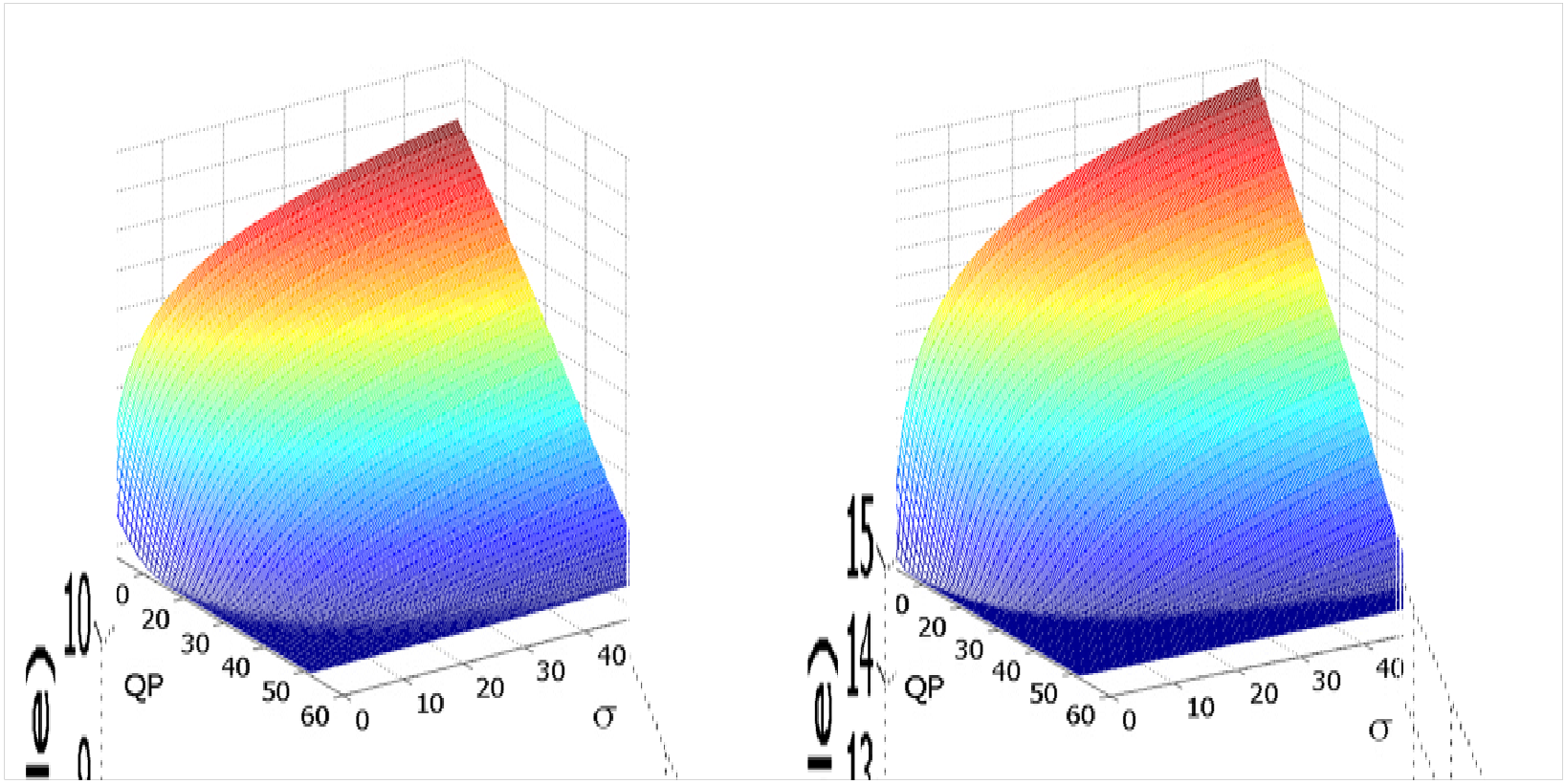
The entropy models determined in the paragraphs 3.2 and 3.3 were compared using 3D graphs. The highest value per sample for Laplace and Gauss models, 8.33 and respectively 8.35 are slightly higher than the number of bits (8) needed to encode the maximum value (255) in the RGB system.

GGD with $\alpha = 1$ confirmed with the value 8.58 the result obtained by the original Laplace distribution (8.33), while the value provided by the entropy model of GGD ($\alpha = 2$) indicated in (Sun and others, 2013b) proved to be too big (14.5 compared to 8 bits/sample). This is because the formula for generalized Gauss, as is provided in (Sun and others, 2013b), was first computed for the Laplace case and then extended to the general case through a variable change. The Figures 3.3 and 3.4 show these discrepancies.



a) b)

Figure 3.3 The entropy (bits/sample) calculated with: a) Laplace pdf and b) Gauss pdf



a) b)

Figure 3.4 The entropy (bits/sample) calculated with GGD: a) $\alpha = 1$ (Laplace) and b) $\alpha = 2$ (Gauss)

CHAPTER 4

MACROBLOCK LEVEL ADAPTIVE LAGRANGE MULTIPLIER COMPUTATION

4.1 Motivation

To recapitulate, there are several ideas related to the value of λ and its usage per frame/macroblock. Firstly, there was the idea of using a constant value for λ, the same for each macroblock, each frame and each sequence, given a QP value. It represents a static approach, where potentially unnecessary bits are transmitted for the same video quality.

A second approach, the state-of-the-art outlined in (Li and others, 2009) and (Li, Oertel and Kaup, 2007), taken over by the articles (Wang and others, 2012) and (Wang and others, 2011), makes the distinction between different sequences and different frames of the same sequence. It aims for a Lagrange multiplier adaptive with the frame content and characteristics, which constitutes a step ahead in bitstream optimization. It has the drawback of considering the macroblocks as being identical from the standpoint of the statistical content. However, macroblocks differ across a frame. They may have different RD curves commanding different a Lagrange multiplier which is a function of the macrobloc characteristics rather than based on a frame level model. We conjecture that if λ had been calculated using macroblock level characteristics, then the gain in terms of PSNR and video quality would prevail over the case when λ is computed at the frame level.

What we would like to get for optimal λ is the value of the slope of the tangent to RD operational curve at the QP we are encoding with, if possible, taking into account macroblock level statistics. Generally, depending on the value of Lagrange multiplier used to weigh in between distortion and rate, there could be two extreme situations. A greater Lagrange multiplier value results in smaller rates, the encoder is biased toward minimizing the rate and the winner mode in the decision process could be Skip mode or, at most, a mode

based on inter- prediction. On the contrary, if the Lagrange multiplier is small, the rate is big, the distortion results in small value, so the encoder is biased toward small distortion and an intra-mode might be declared winner in the mode decision race.

Moreover, since the inter frame prediction of larger blocks is not quite accurate, the allocated bits for the quantized transformed coefficients are in greater number, while those for header are quite a few. Smaller inter coded partitions, much easier to predict, deliver smaller amount of bits per encoded coefficient in contrast with a larger volume for the header that contains the MVs. Intra modes allocate most bits for the quantized transformed coefficients.

In order to assess the percentage of the Gauss/Laplace-type macroblocks several sequences were run with QP = [1...51] and the standard configuration, e.g. UseCustomLM = 1 as described in ANNEX VI. Each macroblock type was assessed using the discrimination criterion described in 2.2.4. The percentage of Gauss macroblocks over the total number was calculated on several frames (1-5) for various types of sequences and displayed in Figure 4.1. The slow-paced sequences (bridge, container) show a saturation of the percentage of Gauss type macroblocks in the first third of the QP range followed by a descending slope, while the medium-paced (foreman, silent) and fast-paced (bus, ice, soccer, coastguard) sequences display maximums. We can observe in some cases that over 70% of the macroblocks have Gauss rather than Laplace distributions. This supports our belief that we should not limit our study to the Laplace distribution.

As QP increases for lower rate, the Laplace-based macroblocks become predominant. Since coding with Laplace has much lower complexity it is preferable to code using greater values of QP. Another strong reason as to why it is preferable to consider macroblock level statistics as opposed to frame level statistics is because Lagrange multiplier computed as a mean value over the whole frame would not be appropriate for all macroblocks, since any macroblock's statistical properties values might be different from the ones of the frame.

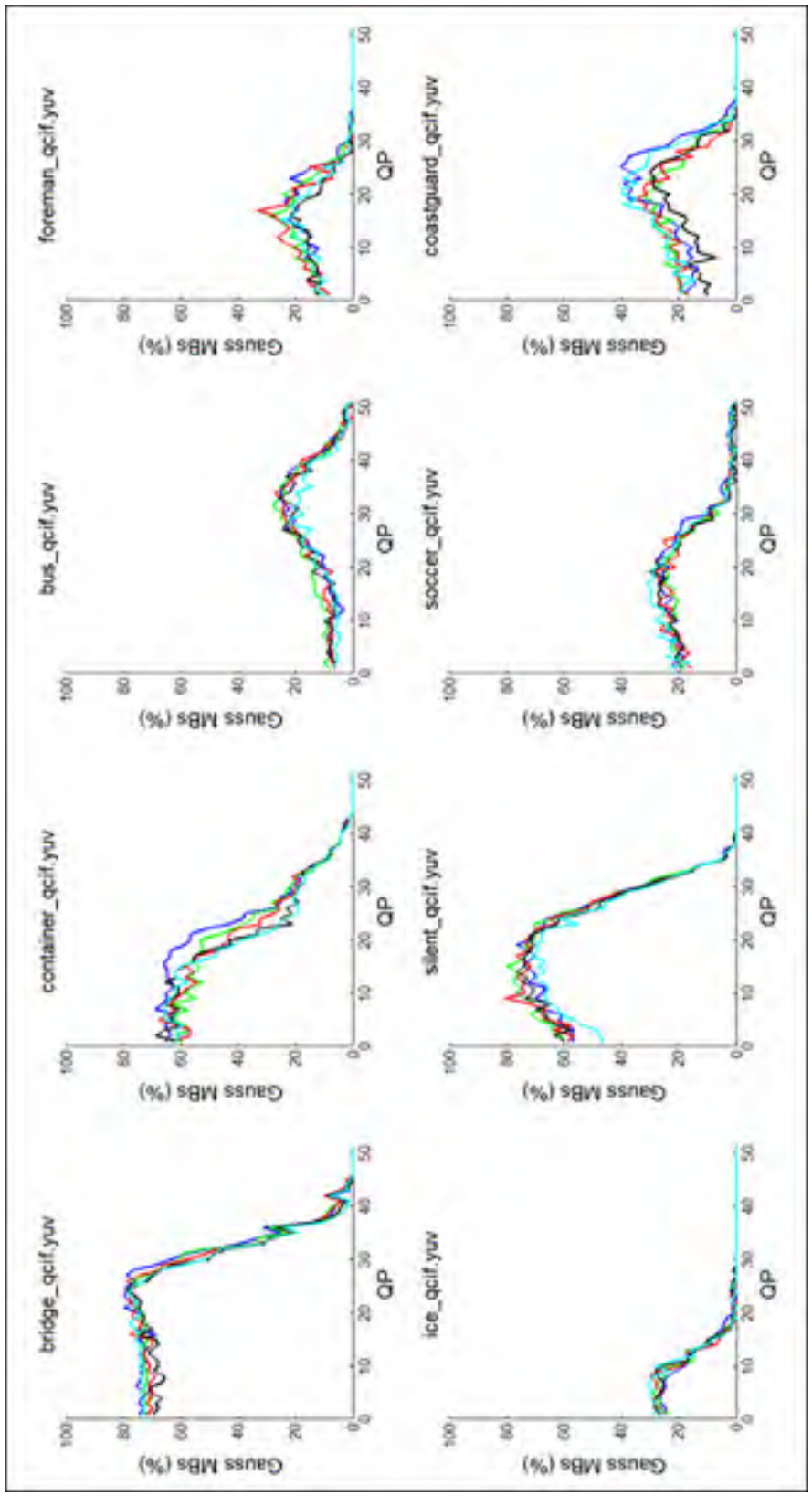


Figure 4.1 The dependence of the percentage of Gauss type macroblocks on QP
a) bridge; b) container; c) bus; d) foreman; e) ice; f) silent; g) soccer; h) coastguard
where frame 1 (blue), frame 2 (green), frame 3 (red), frame 4 (black), frame 5 (cyan)

4.2 RDO using frame level Laplace distribution based Lagrange multiplier

This method has as starting point several experiments conducted at the MB level through which macroblocks are adequately grouped to form entities, named coding units, having the same Laplace parameter. Each coding unit has the same Laplace parameter which may vary from one coding unit to another. The goal is to determine if another frame level model of rate and distortion, better than the ones described in (Li and others, 2009) can be deduced using the MB statistics.

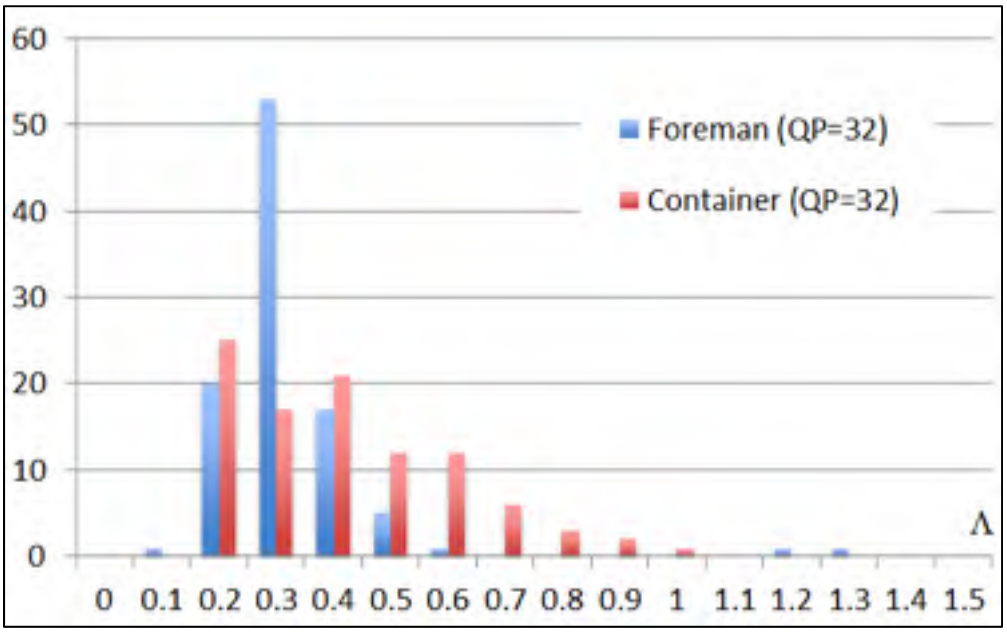


Figure 4.2 Histogram of MB level Laplace parameters for frame 10 of QCIF sequences Foreman and Container encoded at QP = 32

In the first phase of the experiments each coding unit is represented by a single MB and one estimates its Laplace parameter. Figures 4.2 and 4.3 depict the histogram of MB level Laplace parameters for all MBs pertaining to the frame 10 of the QCIF sequences Foreman and Container, encoded in H.264 baseline at QP = 32 and QP = 40 respectively. One can see that the MB level Laplace parameter varies quite significantly within the frame, due to the fact that even for slow motion sequences the macroblocks differ in terms of variance.

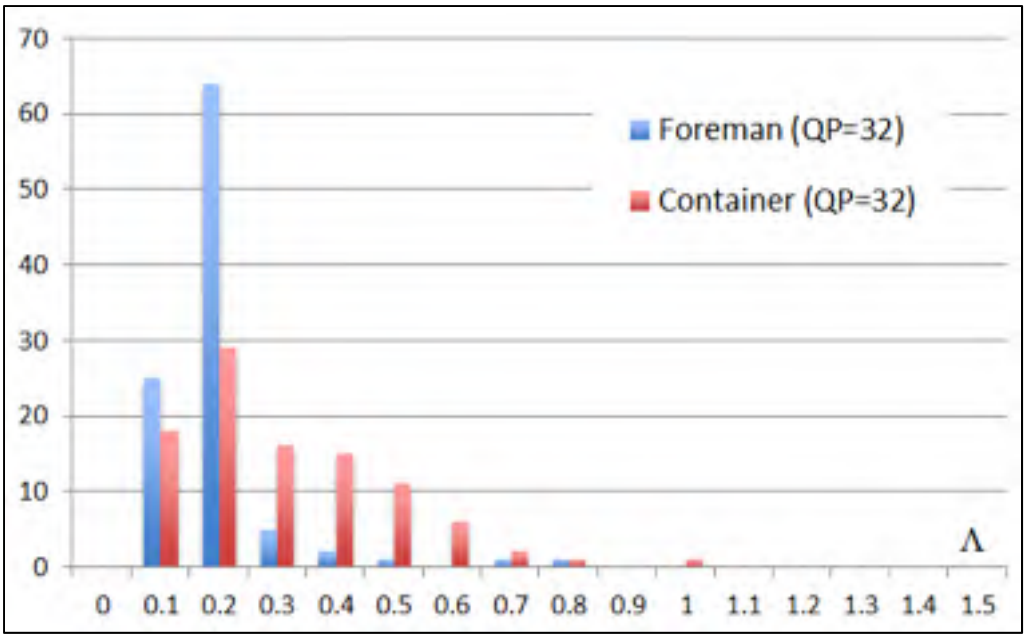


Figure 4.3 Histogram of MB level Laplace parameters for frame 10 of QCIF sequences Foreman and Container encoded at QP = 40

Similar variations have been noticed on other frames on same and different sequences coded at various QPs.

4.2.1 Analysis of multiple coding units having Laplace distribution

We define multiple coding units, for short MCU, as MBs or regions of a frame containing one or more MBs, contiguous or not, having various sizes (the number of MBs can change from one MCU to another), with transformed residuals obeying the Laplace distribution and having distinct Laplace parameters so that they can be easily clustered based on this criterion. Based on this description, we analyze their performance in terms of rate-distortion, mentioning that the results can be further generalized or enhanced based on other supplementary features used as clustering criterion. As each coding unit obeys the Laplace distribution with distinct parameter Λ_i and occupies an area of n_i MBs in the frame, there can be up to K regions in a frame of N MBs and we can calculate the average distortion D and rate R per MB as in (4.1),

$$D = \frac{1}{N}\sum_{i=1}^{K} n_i.D_i \qquad (4.1)$$
$$R = \frac{1}{N}\sum_{i=1}^{K} n_i.R_i$$

where D_i and R_i are respectively the distortion and rate of the i-th coding unit and obviously $\sum_{i=1}^{K} n_i = N$. Using (2.4) we can define the Lagrange multiplier for the MCU model λ_{mcu} as in (4.2).

$$\lambda_{mcu} = -\frac{\sum_{i=1}^{K} n_i \frac{\partial D_i}{\partial Q}}{\sum_{i=1}^{K} n_i \frac{\partial R_i}{\partial Q}} \qquad (4.2)$$

When the regions have the same size in terms of number of MBs, i.e. $n_i = N / K, \forall i$ the expression (4.2) becomes

$$\lambda_{mcu} = -\frac{\sum_{i=1}^{K} \frac{\partial D_i}{\partial Q}}{\sum_{i=1}^{K} \frac{\partial R_i}{\partial Q}} \qquad (4.3)$$

One can see that when all coding units have the same Laplace parameter $\Lambda_i = \Lambda$ then $\lambda_{mcu} = \lambda_{Lap}$ showing that the proposed approach is a generalization of (2.4). According to (2.1), the Laplace parameter Λ_i for the MCU model is estimated in terms of standard deviation σ_i of the transformed residuals of coding unit *i* as:

$$\Lambda_i = \frac{\sqrt{2}}{\sigma_i} \qquad (4.4)$$

where σ_i is the standard deviation of the i-th coding unit. One can easily verify the relationship (4.5) between the variances of the frame σ (considered as a single region) and coding unit levels σ_i.

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{K} n_i.\sigma_i^2 \tag{4.5}$$

It turns out that the Laplace parameter Λ at the frame level can be calculated based on the individual values of the coding units Λ_i , $i = 1...K$ as in (4.6).

$$\Lambda = \frac{\sqrt{2}}{\sigma} = \frac{\sqrt{2}}{\sqrt{\frac{1}{N}\sum_{i=1}^{K} n_i.\sigma_i^2}} = \frac{\sqrt{2}}{\sqrt{\frac{1}{N}\sum_{i=1}^{K} n_i.\frac{2}{\Lambda_i^2}}} = \frac{1}{\sqrt{\frac{1}{N}\sum_{i=1}^{K} \frac{n_i}{\Lambda_i^2}}} \tag{4.6}$$

In order to have an early evaluation in terms of performance of the MCU approach we consider a frame from the QCIF sequence Foreman in Figure 4.3 consisting of three coding units of equal size having the parameters Λ_i = (0.15, 0.25, 0.25). This choice is motivated by the distribution of Λ_i, one third situated roughly in the interval (0.1 - 0.2) and two thirds in (0.2 - 0.3) respectively, as one can see in the picture. For this set of Λ_i the calculated frame level $\Lambda = 0.198$ according to (4.6).

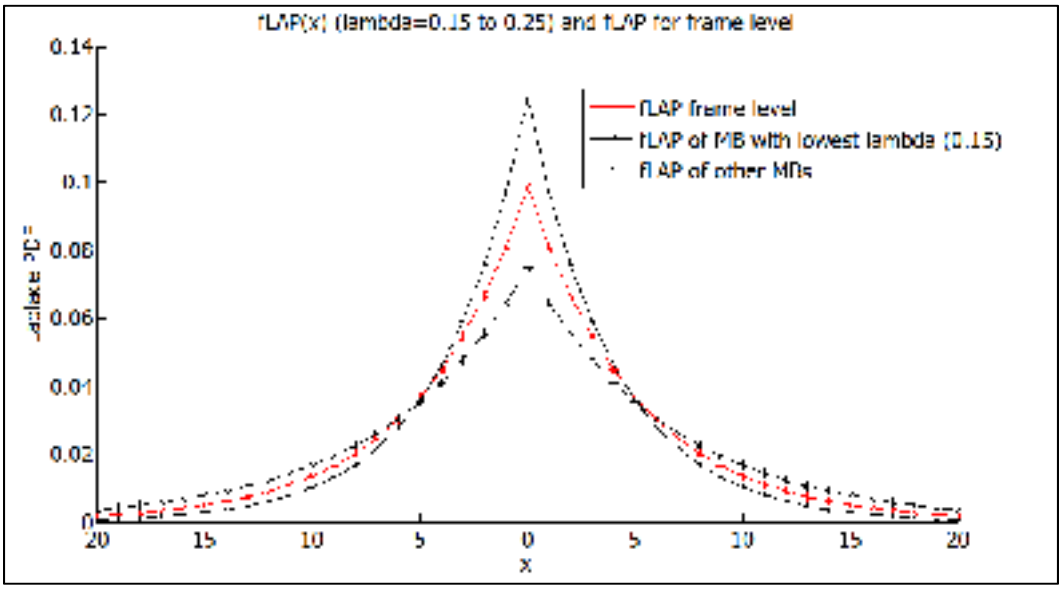


Figure 4.4 Laplace distributions of coding units with $\Lambda_1 = 0.15$, $\Lambda_2 = 0.25$, and $\Lambda_3 = 0.25$ and frame level approximation, for Foreman_qcif.yuv

Figure 4.4 shows the Laplace distributions of the coding units along with the approximated frame level distribution when all MCUs are merged according to (4.6), while Figures 4.5 and 4.6 show the R-D curves and the Lagrange multipliers λ_{Lap} and λ_{mcu} with respect to λ_{HR}, used in the H.264 Joint Model (JM). We notice that a very small difference between the Λ_i compared to the span of Laplace parameter values observed in Figures 4.2 and 4.3 entails a difference between the R-D curves and, what is more important, in the set of computed Lagrangian multipliers. Figure 4.5 shows that it is more advantageous to use a Lagrange multiplier based on the relation (4.2) than compute one at a frame level based on the merged regions. In order to evaluate the effect of merging the regions according to (4.6) and experiment with the Lagrangian multiplier based on the relation (4.2), we used a set of possible values for the Lagrange multiplier, up to 10^{30}.

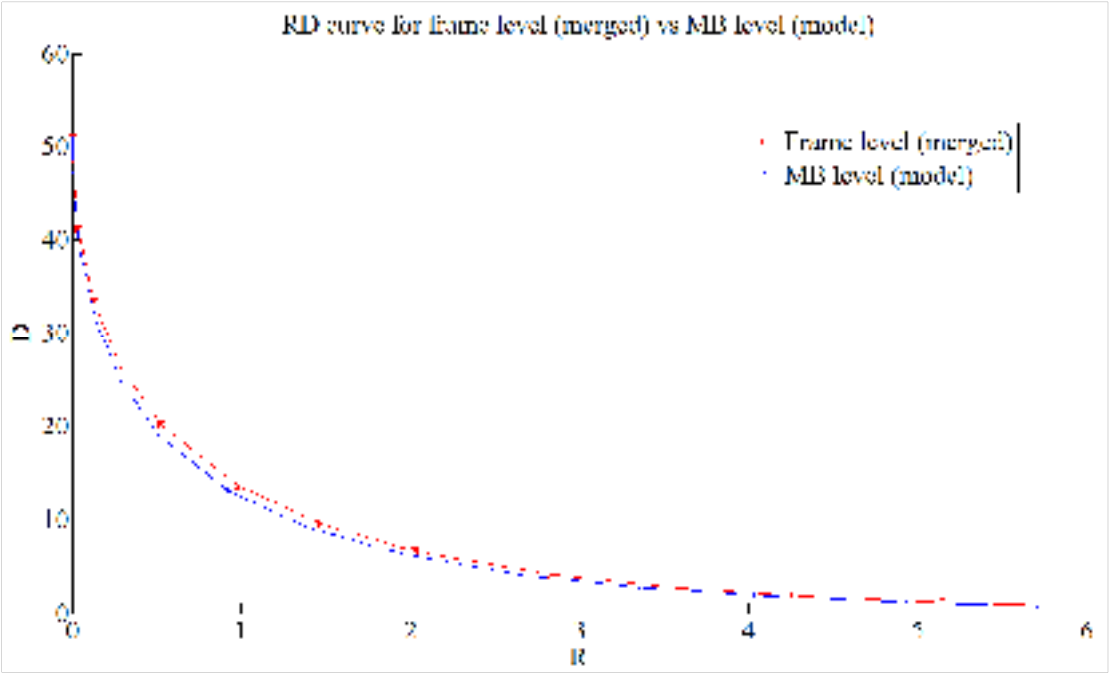


Figure 4.5 R-D curve of MCU and frame level approximation for MCU comprised of $\Lambda_1 = 0.15$, $\Lambda_2 = 0.25$, and $\Lambda_3 = 0.25$, for Foreman_qcif.yuv

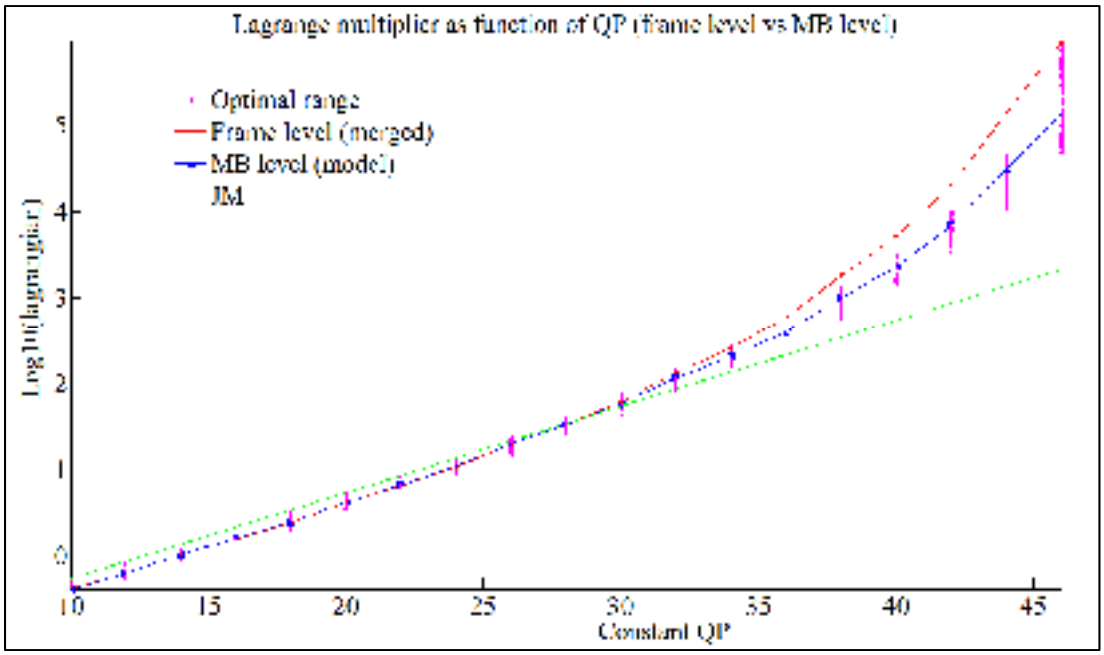


Figure 4.6 Lagrange multiplier for MCU and frame level approximation for MCU comprised of $\Lambda_1 = 0.15$, $\Lambda_2 = 0.25$, and $\Lambda_3 = 0.25$, for Foreman_qcif.yuv

Interesting conclusions can be drawn from the analysis of Figure 4.6: for each value of the Lagrange multiplier in the designated set, there is an optimal value of the QP (the optimal range depicted in pink dots). Conversely, each QP is associated a range of optimal Lagrangian multipliers. We notice that the λ_{mcu} values calculated using (4.2) situate within the optimal range at each QP. As a matter of fact, starting from around QP = 28, $\lambda_{HR} < \lambda_{mcu} < \lambda_{Lap}$. However, at high QPs, the frame level Lagrange multiplier value based on (4.6) fails to reside within the optimal range, so for high QPs we should code using a Lagrangian based on (4.2) formula.

We drew the same conclusions from other experiments with different sequences, MCU configurations, and Laplace parameters. For example, Figure 4.7 shows the Laplace distributions of the coding units with $\Lambda_1 = 0.25$, $\Lambda_2 = 0.35$, $\Lambda_3 = 0.45$, $\Lambda_4 = 0.55$, and $\Lambda_5 = 0.65$, along with the approximated frame level distribution when all MCUs are merged according to (4.6). The RD curves and Lagrange multipliers ($\lambda_{Lap}, \lambda_{mcu}$) are depicted in Figure 4.8 and 4.9 respectively.

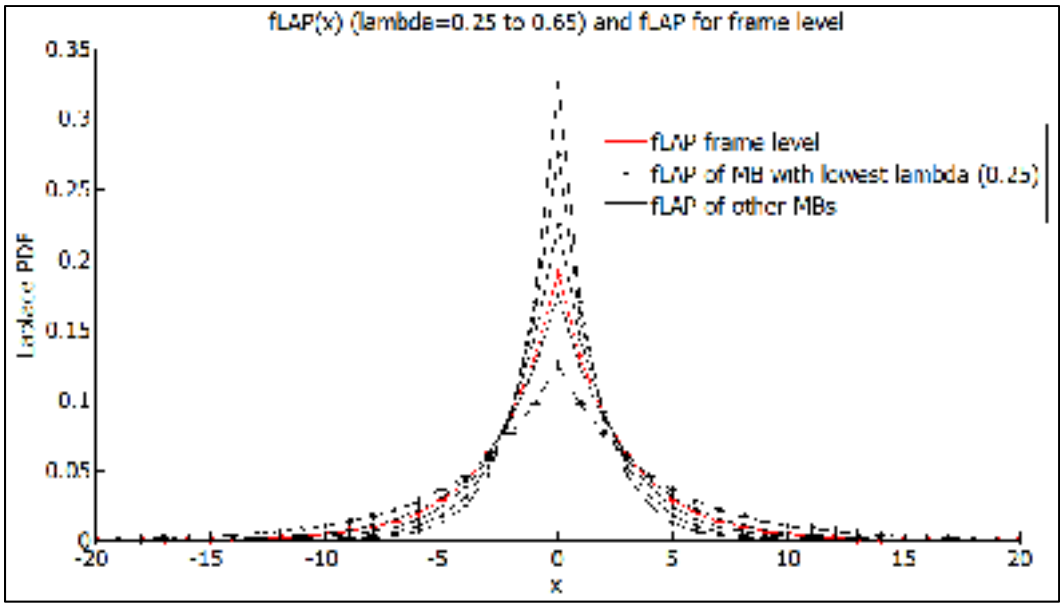


Figure 4.7 Laplace distributions of coding units and frame level approximation for $\Lambda_1 = 0.25$, $\Lambda_2 = 0.35$, $\Lambda_3 = 0.45$, $\Lambda_4 = 0.55$, and $\Lambda_5 = 0.65$, for Container_qcif.yuv

This configuration represents a distribution of MBs related to the sequence Container in Figure 4.2, which is similar to the previous analyzed distribution. Now, the differences between R-D curves and Lagrange multipliers have become larger.

So far, based on the assumption that the regions obey the Laplace distribution with different Laplace parameters, the experiments showed that only a Lagrangian multiplier based on the regional RD models might be optimal. In this case we need to calculate the derivatives of rate and distortion at the region level with (4.2). A faster frame level approach, based on (4.6), might be acceptable only when the regions have similar Laplace parameters. In practice, as we will show in Figure 4.9, depending on QP and not only on it, a MB transformed residuals setting might fit the Laplace, Gauss or mixtures of those.

These conclusions were drawn for the case when all regions obey Laplace distribution, but they might be extended and enriched while experimenting with other distributions, such as Gauss, or combinations of them. For those cases, appropriate models must to be put in place.

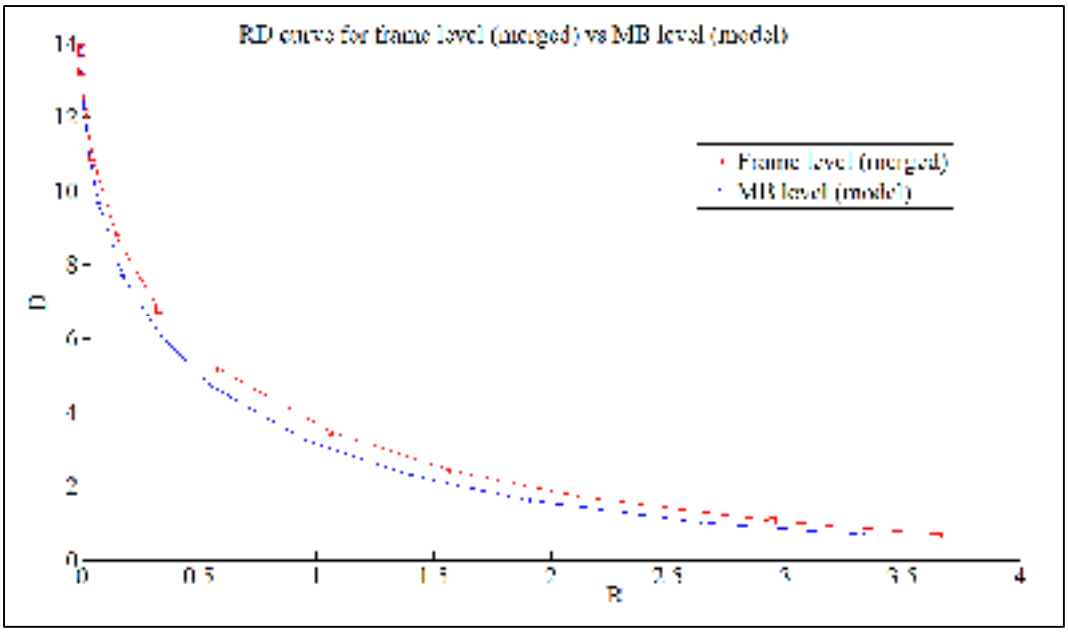


Figure 4.8 R-D curve of MCU and frame level approximation for MCU comprised of $\Lambda_1 = 0.25$, $\Lambda_2 = 0.35$, $\Lambda_3 = 0.45$, $\Lambda_4 = 0.55$, and $\Lambda_5 = 0.65$, for Container_qcif.yuv

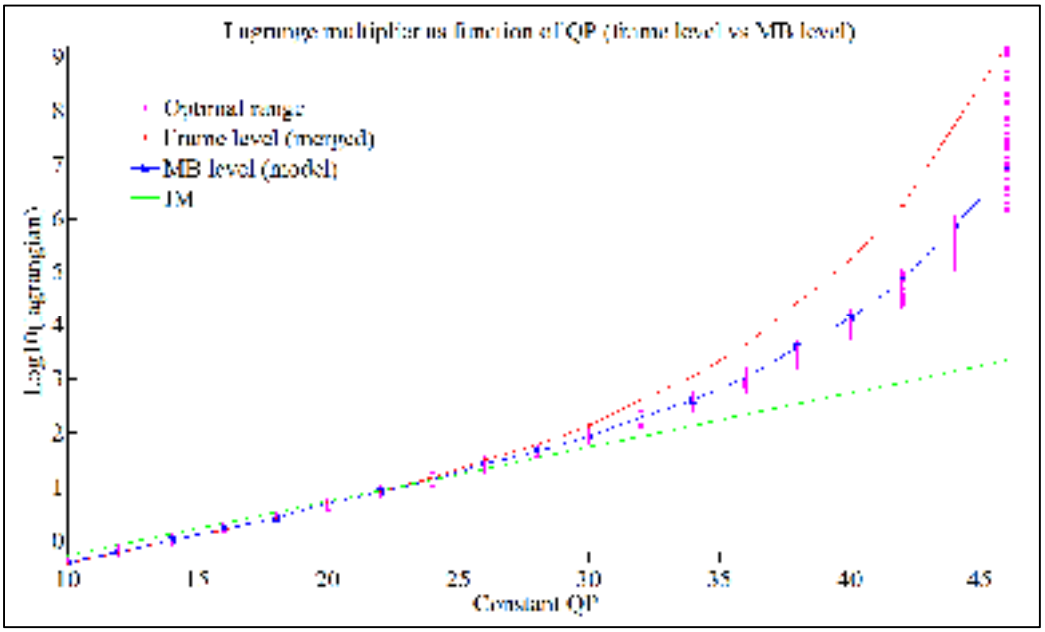


Figure 4.9 Lagrange multiplier for MCU and frame level approximation for MCU comprised of $\Lambda_1 = 0.25$, $\Lambda_2 = 0.35$, $\Lambda_3 = 0.45$, $\Lambda_4 = 0.55$, and $\Lambda_5 = 0.65$, for Container_qcif.yuv

As MCU-based method seems to be the approach of choice, the problem of predicting the region size arises, namely if the regions are too small then the error of predicting the distribution parameters increases too much, while larger regions would face the issue of mixing distributions of same type or different types. For the case when the prediction error of the distribution parameters is too high, a frame level Lagrangian based on (4.6) might prove a better choice, but this case should be carefully analyzed, since a high QP might have an undesired impact. In this work we considered MCU size limited to a single MB. Under this assumption, a set of 256 transformed residuals coefficients is expected to be large enough to reasonably estimate the parameters of any possible distribution under consideration, including Laplace and Gauss and their mixtures.

4.2.2 Macroblock level processing

Firstly, the processing at MB level has to take into consideration the type of the distribution that closest fits the transformed residual coefficients. We previously calculated the expressions of R_i and D_i functions for several distributions such as Laplace (3.6 and 3.7), and Gauss (3.9 and 3.10). From (Sun and others, 2013b), we have these expressions in the case of generalized Gauss (3.12 and 3.13). We have also computed the derivatives of R_i and D_i with respect to Q (see ANNEX I, ANNEX II, and ANNEX III).

For a skipped MB, we determined the distortion expressions namely (3.8) for Laplace distribution and (3.11) for Gauss distribution respectively. In this case a single bit is transmitted and both the derivatives of distortion and rate are considered equal to zero, since their rate and distortion values do not depend on Q.

$$\partial D_i/\partial Q = 0 \qquad (4.7)$$
$$\partial R_i/\partial Q = 0$$

We want to compare this algorithm with to the state of the art (Li and others, 2009) to see where we stand in terms of entropy. Let N be the number of MBs in the frame and $N_{i,n}$, the

number of transformed coefficients in the $i-th$ MB mapped to quantization level n. The total number of transformed coefficients in the frame can be calculated in two ways:

$$\sum_{i=1}^{N}\sum_{n=-\infty}^{\infty} N_{i,n} = 256 \times N \tag{4.8}$$

When all MB are treated equally, without the possibility that actually the MB be skipped, at the frame level the initial probability associated to the $n-th$ quantization level is:

$$P_n = \frac{\sum_{i=1}^{N} N_{i,n}}{N \times 256}, n \in Z \tag{4.9}$$

Of all N MBs, let W be the number of MBs that are coded as skipped. It follows that P_n adjusts to the following expression, noted as P_n^*, which excludes the transformed coefficients pertaining to the set S of skipped MBs:

$$P_n^* = \frac{\sum_{i=1, i \notin S}^{N} N_{i,n}}{(N-W) \times 256}, n \in Z \tag{4.10}$$

The authors of (Li and others, 2009) assumed that all transformed coefficients within skipped MBs are quantized to level 0. In this case, one can define N_0, the total number of transformed coefficients quantized to level zero (including those belonging to skipped MBs) and N_S, the total number of transformed coefficients that belong to the set S of skipped MBs only:

$$N_0 = \sum_{i=1}^{N} N_{i,0}$$
$$N_S = \sum_{i=1, i \in S}^{N} N_{i,0} \tag{4.11}$$

Thus, the probabilities P_0 of the transformed coefficients quantized to bin zero over all MBs (including those in skipped MBs), and P_S of the skipped MBs, become:

$$P_0 = \frac{\sum_{i=1}^{N} N_{i,0}}{N \times 256} = \frac{N_0}{N \times 256} \qquad (4.12)$$
$$P_S = \frac{\sum_{i=1, i \in S}^{N} N_{i,0}}{N \times 256} = \frac{N_S}{N \times 256}$$

The probability P_0^* that excludes the skipped MBs becomes:

$$P_0^* = \frac{N_0 - N_S}{(N-W) \times 256} = \frac{N_0 - N_S}{(1 - W/N) \times N \times 256} = \frac{P_0 - P_S}{1 - P_S} \qquad (4.13)$$

It follows that:

$$P_n^* = \frac{P_n}{1 - P_S} \qquad (4.14)$$

So the average entropy per transformed coefficient (we assume skipped blocs do not affect entropy):

$$H^* = \frac{N-W}{N}\left(-P_0^* . \log_2 P_0^* - 2\sum_{n=1}^{+\infty} P_n^* . \log_2 P_n^* \right) \qquad (4.15)$$
$$= (1 - P_S)\left(-P_0^* . \log_2 P_0^* - 2\sum_{n=1}^{+\infty} P_n^* . \log_2 P_n^* \right)$$

With this, it has been demonstrated that this new approach that computes the pair rate-distortion at the MB level without taking in consideration the r factor but assigning zero

value to the derivatives of rate and distortion when the MB is skipped is compatible with the work of (Li and others, 2009) and might exceed their results.

4.3 Macroblock level adaptive Lagrangian multiplier computation

Following the cost function minimization (1.26), the Lagrange multiplier can be computed in several ways, depending on the way the distortion and rate are expressed. The general formula of the Lagrange multiplier is:

$$\lambda = -\frac{dD}{dR} \tag{4.16}$$

When distortion and rate have analytical forms that depend on multiple variables, some of which are described in 1.3.5. and 1.3.6., it is difficult to derive according to (4.1). For this reason, the derivation is made with respect to Qstep.

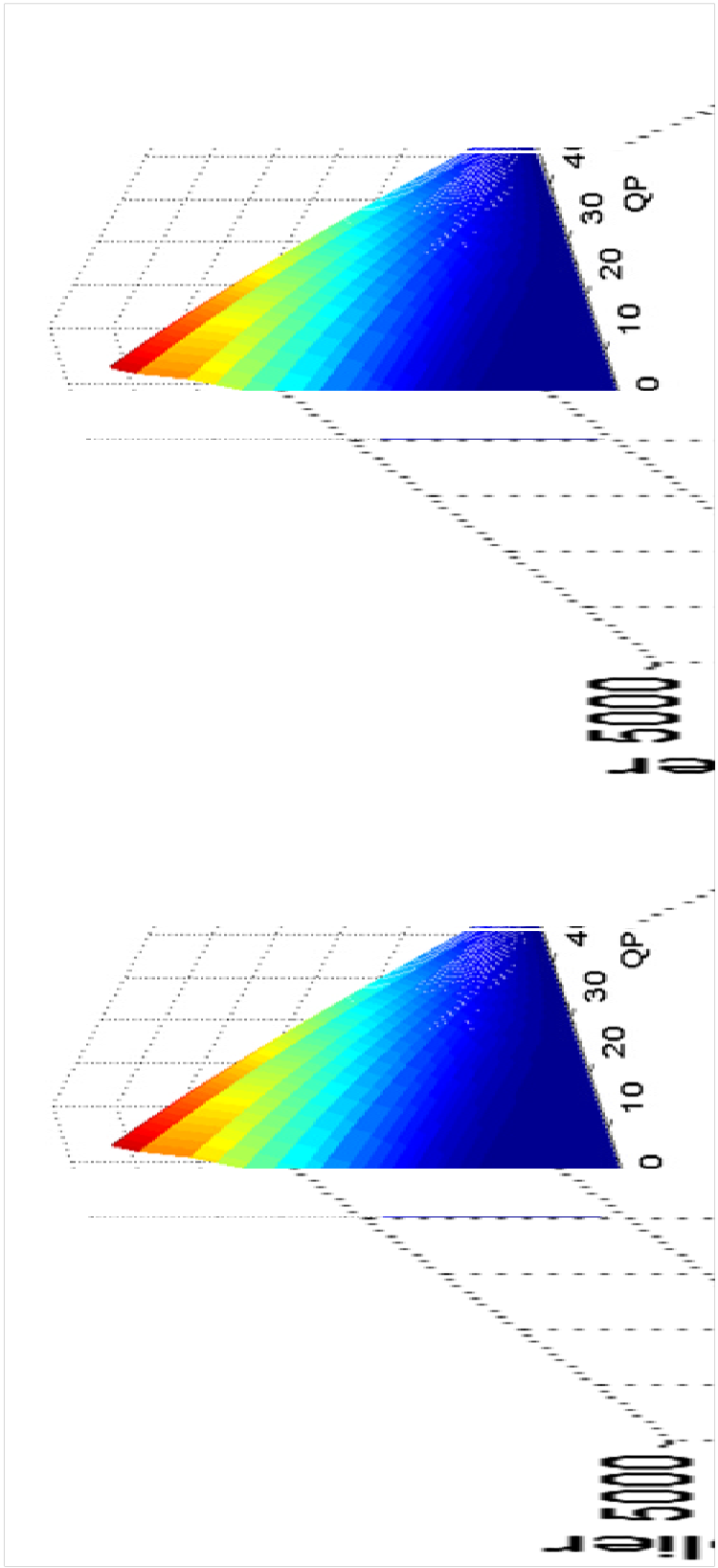
$$\lambda = \frac{-\left(\frac{\partial D}{\partial Q}\right)}{\left(\frac{\partial R}{\partial Q}\right)} = \frac{-\left(\sum_i \frac{\partial D_i}{\partial Q}\right)}{\left(\sum_i \frac{\partial R_i}{\partial Q}\right)} \tag{4.17}$$

For the case of generic model, we calculate the value of λ at discrete points QP.

$$\lambda = -\frac{\Delta D}{\Delta R} = -\frac{\sum_i \Delta D_i}{\sum_i \Delta R_i} \tag{4.18}$$

4.3.1 The optimal Lagrange multiplier as a function of QP for Laplace distribution based model

Laplace-based Lagrange multiplier calculated with the formula (see ANNEX I) ranges from 0 to 4390, a domain that is pretty close to the one (0.07- 4379) generated from a generalized Gauss distribution with $\alpha = 1$ (ANNEX III), but much less than the maximum value (6963) of λ_{HR}. It can be seen that up to QP = 28-30, the value of λ is very low, after which a sharp, exponential steep is recorded, as illustrated in Figure 4.10. The region where λ values are significant is an indicator to consider encoding with those values in order to save the bitrate.



a) b)

Figure 4.10 Lagrange multiplier: a) Laplace-based; b) GGD-based ($\alpha = 1$)

4.3.2 The optimal Lagrange multiplier as a function of QP for Gauss distribution based model

The maximum Lagrange multiplier attainable with Gauss probability distribution function is 3239, much smaller than that of λ_{HR}, 6963. Using the distortion and rate formulae as mentioned in (Sun and others, 2013b), whose derivative with respect to QP(Qstep) are considered in the calculation of λ, a huge value of λ (27816) was found. A 3D representation, as in Figure 4.11, showed that λ would only depend on QP (Qstep) and practically is independent of β (which is equal to σ). The huge value comes from the problem signaled in paragraph 3.7.

4.3.3 The optimal Lagrange multiplier for generic distribution based model

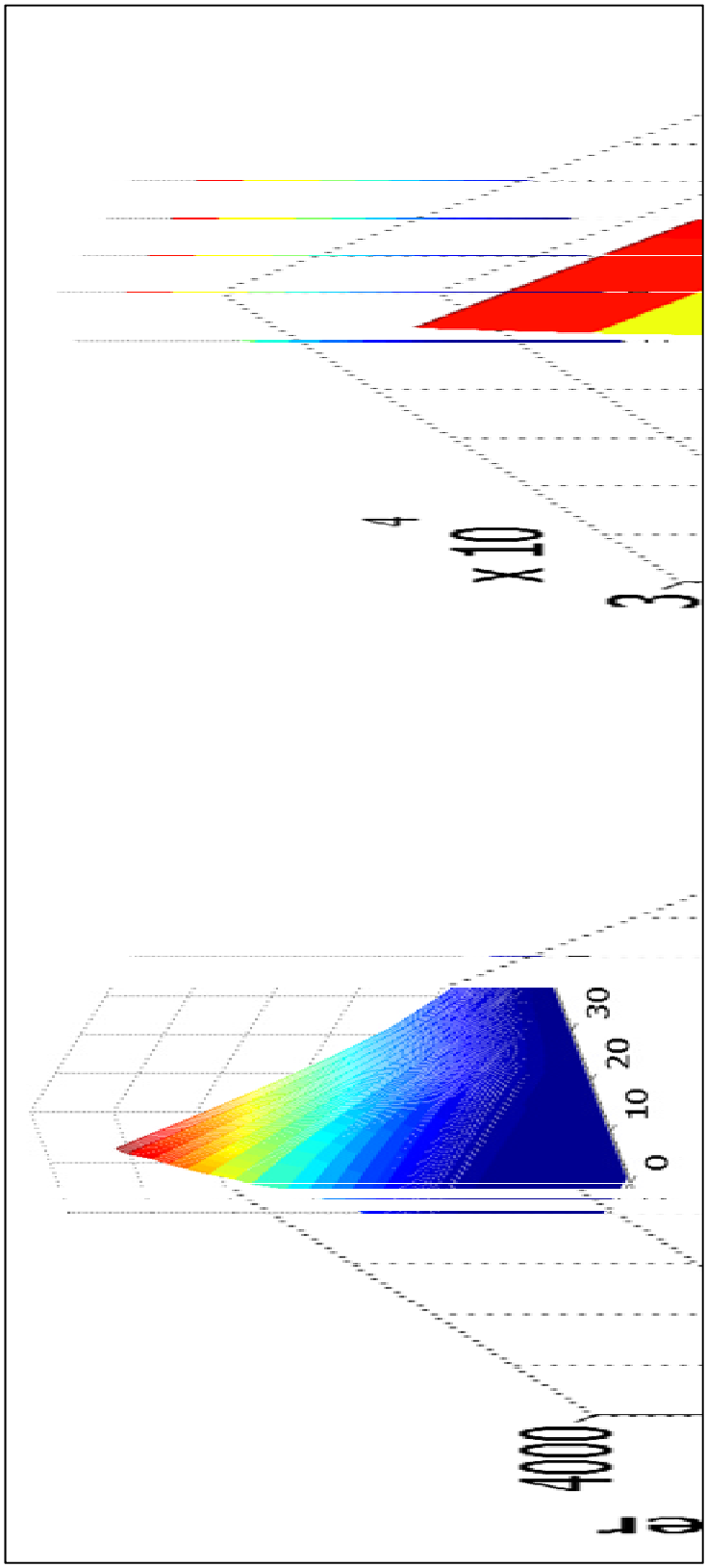
For the model based on numerical integration, the Lagrangian multiplier was computed the following formula:

$$\lambda_{Int} = -\frac{\Delta D_{Int}}{\Delta R_{Int}} \tag{4.19}$$

The discrete value of λ_{Int} was approximated by varying the Qstep with a small offset $\Delta QP = \pm(1-3)QP$ around the central value QP of the current encoding.

$$\lambda_{Int} = -\frac{D_{Int}(QP-\Delta QP) - D_{Int}(QP+\Delta QP)}{R_{Int}(QP-\Delta QP) - R_{Int}(QP+\Delta QP)} \tag{4.20}$$

The offset is chosen so that the video quality would not change drastically. Usually, the operational curves D=D(R) display irregularities as in Figure 4.12, that translate into slope sign changes.



a) b)

Figure 4.11 Lagrange multiplier: a) Gauss-based; b) GGD-based ($\alpha = 2$).

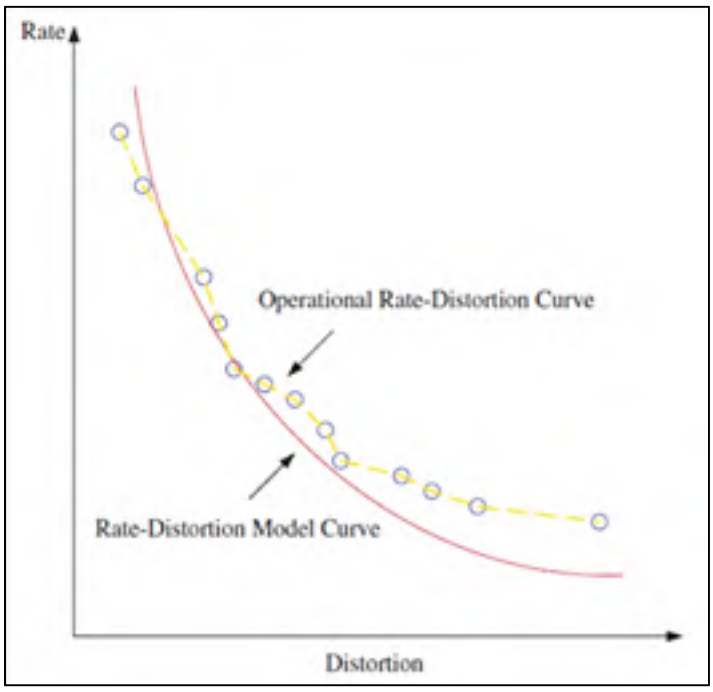


Figure 4.12 Operational RD and RD model curves.
Adapted from (Chen and Ngan, 2007)

Also, the difference $R_{Int}(QP-\Delta QP)-R_{Int}(QP+\Delta QP)$ should not be higher than an acceptable threshold. The iterative process of finding λ_{Int} skips the faulty values and proceeds until a valid one is retrieved. Otherwise, it adopts the static λ_{HR} as a last resort.

4.4 Rate distortion optimization using the macroblock level adaptive Lagrange multiplier computation applied to H.264 compression

When it comes to finding the best prediction followed by the best mode of compression for a macroblock, the Lagrange multiplier comes into play by trading the influence of distortion term over the allocated rate in the cost function expression. As long as λ is based on the QP value only, the problem of deciding its share of rate in the final bitstream is completely solved out. Things get more complicated when the macroblock statistical characteristics need to be considered as part of the λ calculation, as it is intended in the present research.

With the new approach the value of the Lagrange multiplier λ cannot be calculated before knowing the residuals. On the other hand, the residuals of the current macroblock are only available upon encoding with a beforehand known λ since it influences the mode selection.

The solution to this dilemma can come from either using a value of λ determined in the previous frame or macroblock or encoding in two steps with several variants of the latter. Secondly, Lagrange multiplier must be available even before deciding the modes, in order to determine the MVs of the motion compensated prediction with respect to the aimed accuracy (FPel, HPel or QPel). The MVs are determined using a value of λ_{MOTION} as in the equation (1.4).

After the ME process has completed the associated distortion is not retained for further evaluation of the performance, but its resulting rate (number of bits to encode the motion vectors, the prediction method) becomes part of the final bitstream necessary to encode that macroblock, and, in the end, the whole frame. After all, the mode selection is the one that determines the final encoding, that is why the Lagrange multiplier for mode decision overweighs in importance the one associated to motion estimation.

Besides, one needs to select QP values around which the compression is performed. The approach that keeps the same QP along the macroblock/frame encoding with small variations in order to discern the most appropriate value of λ has been adopted.

The one step encoding diagram in Figure 4.13 presents the algorithm that uses the values of distortion, rate and Lagrange multiplier calculated in the previous frame (one frame delayed method) to encode the current one.

A two-step encoding method would first acquire the transformed residuals necessary to estimate the Lagrange multiplier and would encode the macroblock in the second step using that value. Obviously this approach needs significantly more computational resources than the current JM implementation to achieve the compression.

4.5 Summary of the methodology used for experiments

A MB can be encoded using the Laplace model only, in which case the criterion (2.32) is neglected. Otherwise, depending on the distribution type of its transformed residuals, it may be encoded as either Laplace of Gauss. For each MB_i that is not skipped, we compute D_i and R_i as in (AI.13) and (AI.15) for Laplace type or (AII.28) and (AII.5) in the case of Gauss distribution.

The derivatives of D_i and R_i are computed using (AI.16) and (AI.17) in the case of Laplace distribution or (AII.30) and (AII.6) when the distribution is of type Gauss. If the MB is skipped, we use for the distortion calculation either the equation (AI.14) for Laplace type, or (AII.45) for Gauss type. In both cases the rate counts as 1 bit/sample while both derivatives count as zero values. A λ at the frame level is calculated with (4.3) and is used to encode each MB in the next frame. Alternatively, either $\lambda_{LAPLACE}$ or λ_{GAUSS} can be calculated respectively with (AI.18) or (AII.46) while λ at frame level would result by applying (4.6).

Unlike the Figure 1.8, the proposed algorithm connected the block that generates the Lagrangian multipliers (in grey) to the transform block, which provides the samples distributed according to Laplace/Gauss pdfs. Additionally, the block contains an evaluation of the MB statistical parameters, necessary to compute the Lagrangian multipliers.

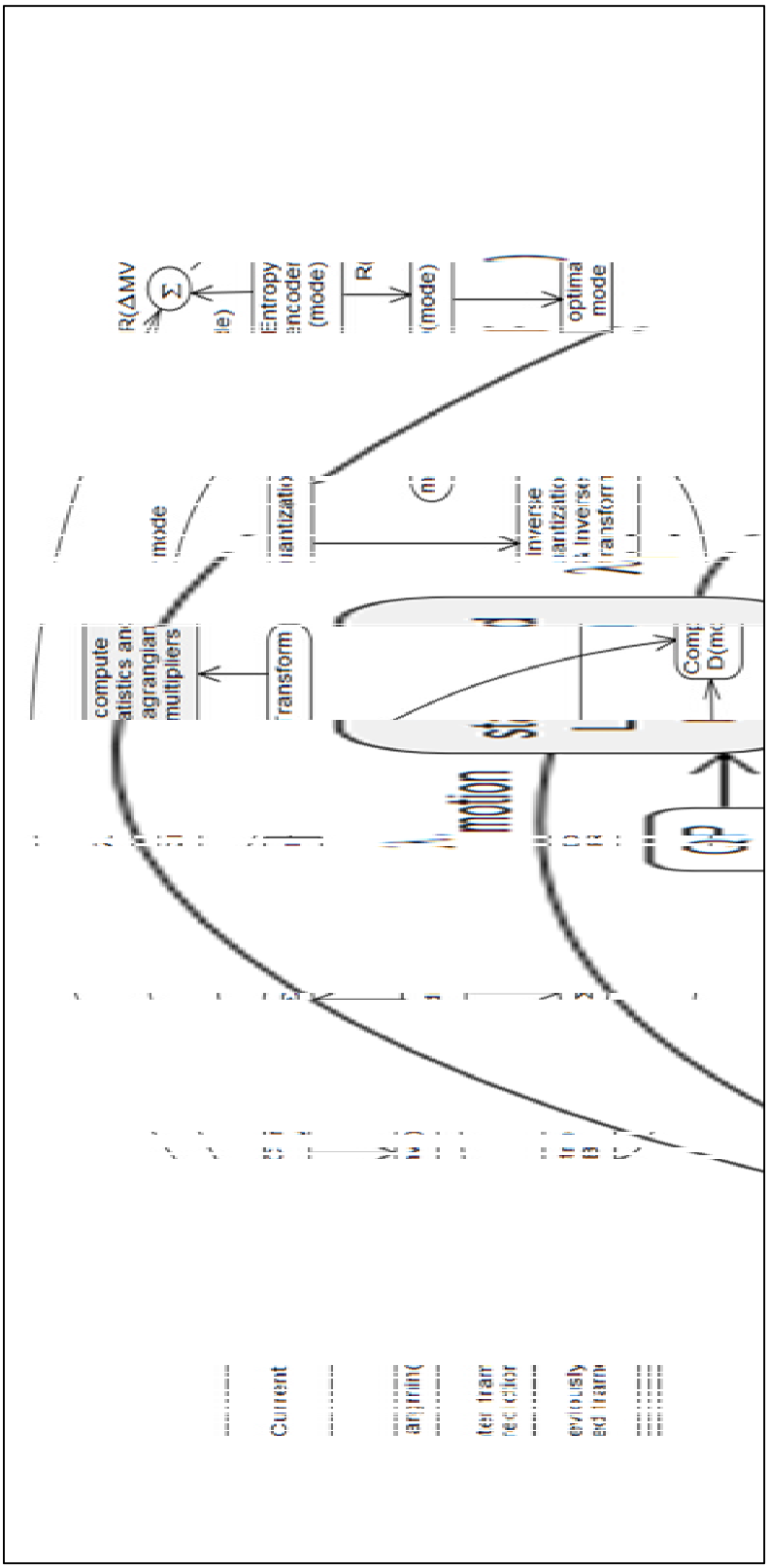


Figure 4.13 Block diagram of one step and one frame delay encoding.
Adapted from (Richardson, 2010)

CHAPTER 5

EXPERIMENTAL RESULTS

5.1 Experimental setup

The experiments were made using a computer Dell XPS 8500 powered by 8 CPUs i7-3770 at 3.40GHz, 12GB installed RAM, and running Windows 7 Enterprise 64-bit.

The experimental environment includes a set of twelve video sequences in format QCIF and CIF of different types of motion (YUV test sequences, 2010):

- slow-paced: container, tempete.
- portrait-type with quasi-static background: foreman, silent.
- fast-paced: bus, coastguard, football, ice, mobile, soccer.

The ANNEX VI contains the configuration file used at the encoding of the sequence carphone_qcif.yuv. All the sequences were rendered using this file adapted according to the utilized method and video sequence.

The field UseCustomLM was added to the configuration file, whose values reflect the encoding method utilized at the macroblock level: Laplace, Gauss, mixed (Laplace/Gauss), numerical integration, and generalized Gauss. The results are compared to the standard approach and the state-of-the-art - adaptive Lagrange multiplier at frame level based on Laplace distribution - described in (Li and others, 2009) and (Wang and others, 2012).

JM version 18.3 was run on each sequence of 100 frames formatted as 1I+99P, with five reference frames and fixed quantization parameter QP = (24, 28, 32, 36, 40, 44), the same for intra as for inter frame. These values were used along with the algorithm to compute the BDPSNR and BDRATE metrics as recommended in (Bjontegaard, 2001). As in (Li and others, 2009), the first intra frame is not considered in the rate-distortion performance analysis. The adaptive rounding and deblocking filters were disabled. RD optimization was

set to high complexity mode to take advantage of the various partitioning modes of the macroblock. The first intra frame is not considered in the encoding using the new approach. The first inter frame is encoded using λ_{HR} and is only used to initialize statistics. The inherent property r of the input sequence, primarily utilized in (Li and others, 2009) is not employed here. Also, the refresh algorithm specific to (Li and others, 2009) was not utilized.

The constants, whose meaning was mentioned in the algorithm of (Li and others, 2009) were utilized with the same values: ζ = 0.35 (for CAVLC), S = 1.982 (for inter frame, either Laplace or Gauss), γ = 1/6 (for inter frames).

For the encoding of the new approach that calculates λ_{mcu} we have used two models: either standalone Laplace or the mixed model Laplace-Gauss. Intra coding in interframe and PCM were disabled. We chose to simulate only these two models for modifying the Lagrange multiplier because they are the only two susceptible to be implemented in a final product and are numerically stable for computing the Lagrange multiplier. Nevertheless, rate and distortion estimation performance was performed on each of the models studied in this research work mostly for comparison and cross-validation purposes (e.g. to make sure that our Laplace and Gaussian models were well implemented and accurate enough).

For a non-skipped MB the distortion and rate derivatives are used to calculate the Lagrange multiplier. On the contrary, for the skipped ones, the distortion and rate derivatives are deemed equal to zero.

The Lagrange multiplier is calculated over all MBs in the frame. In each frame the current value of the Lagrange multiplier λ is calculated based on the current values of the distortion and rate derivatives, which are based on the case of the Laplace distribution on the current value of the Laplace parameter Λ .

Several levels of saturation were imposed: (0.9 ... 5.0) * λ_{HR} and (0.9 ... 5.0) * $\lambda_{previous}$ whenever the computed value of λ was too low or respectively too high. A mean of λ values

spanning the last five frames is performed to compute the final Lagrange multiplier of each frame.

The search method for prediction was set to EPZS, the motion vectors were detected at FPel level using SAD metric, while for the accuracy levels of HPel and Qpel, SATD was employed. The modes were decided using the SATD metric.

5.2 Model parameters estimation at the macroblock level

This research firstly focused on modeling the DCT residuals with the probability distribution functions of either Laplace or Gauss, considered separately, since at the macroblock level the distribution of transform residuals might have been close to either one of them, as opposed to the Laplace-type frame level.

For the cases where neither Laplace nor Gauss pdf is applicable entirely over the whole MBs in the frame, a decision as to what distribution is a better fit for the residuals' real shape was made by using a goodness of fit test, outlined in paragraph 2.2.4.

A fourth way uses the generalized Gauss distribution, which covers symmetric Laplace, Gauss, and uniform distributions. In this case, the statistical test to discriminate between Laplace and Gauss is irrelevant because, based on the parameters value, the distortion and rate are calculated using the formulae in (Sun and others, 2013b).

A more general approach based on numerical integration, independent of the pdf shape, skewness, and kurtosis has completed the analysis. It covers any type of shape of the real pdf, discarding all assumptions previously made such as in the case of Laplace pdf, when the integration was made with the assumption the samples set is zero-mean (mean = location = 0), which is true at the frame level, but it does not always occur at the macroblock level, due to the insufficient number of samples. Using these five models at macroblock level, the

distortion, rate and Lagrange multiplier were calculated, tested and compared to JM, the standard implementation of H.264.

Further, using Matlab, we developed a helper application which receives as input data the transformed residuals at the frame/MB levels and plots the whole map of the probabilities of distribution functions at the frame/MB levels. These plots have shown a good fit of the transformed residuals with the Laplace or Gauss pdf, discriminated with the criterion 2.2.4.
This application helped to better understand why one can rely on the distribution type (Laplace) of the transformed residuals at the frame level, and why at the MB level, it is more difficult to predict the distribution type, mainly due to the lack of a sufficient number of samples to define the distribution parameters.

For example, the distribution of transformed residuals was analyzed at the frame level for two values of QP, 20 and respectively 36, as in the Figure 5.1 a) and b). In Figure 5.1, it is depicted the real distribution of the transformed residuals (*Tr.Res.*) along with the theoretical distributions of the Laplace model (*Laplace*) and Gauss model (*Gauss*) that have the same characteristics, (η, θ) respective (μ, σ) as the real transformed residuals.

While at the frame level the transformed coefficients' distribution is clearly of Laplace type (frame 2(P) of Figure 5.1) no matter the value of the quantization parameter, at the macroblock level, the distribution type depends on QP. The sequence container was rendered with mixed model that selects the appropriate distribution, Laplace or Gauss, based on the discrimination criterion described in 2.2.4.

For QP = 20, ten macroblocks ([7,1], [7,2], [7,3], [8,1], [8,2], [8,3], [9,1], [9,2], [9,3], [9,4]) have fit the Gauss distribution as shown in Figure 5.2, while for QP=36, the same macroblocks in the same frame have their transformed coefficients distributed according to the Laplace distribution, as illustrated in Figure 5.3.

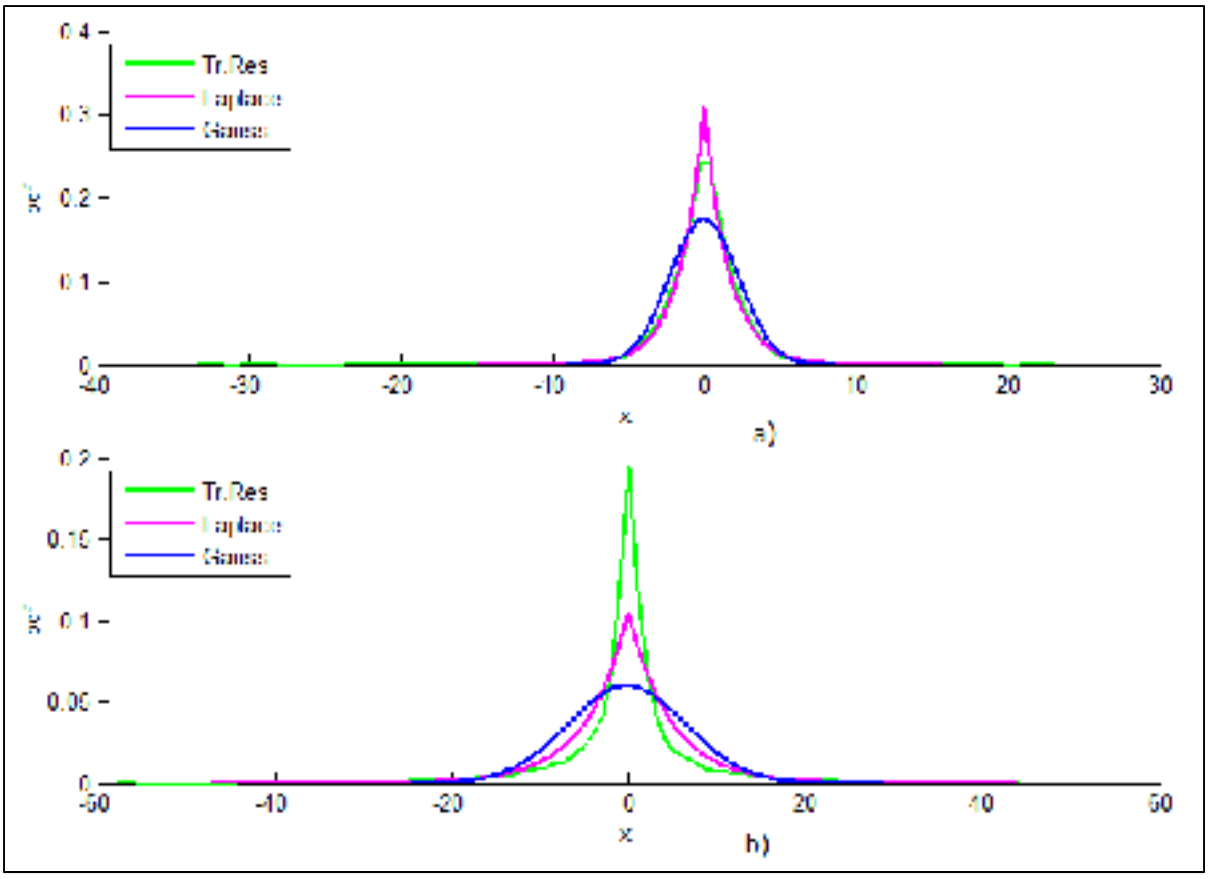


Figure 5.1 Distribution of frame transformed residuals (container, frame 2(P)), a) QP = 20; b) QP = 36

This confirms the conclusions related to Figure 4.1 and shows that one can estimate well the parameters related to both distributions.

The problem is that the number of samples that is available for each MB is only 256, which now seems insufficient to determine the exact parameters and even then nature of the distribution of residuals.

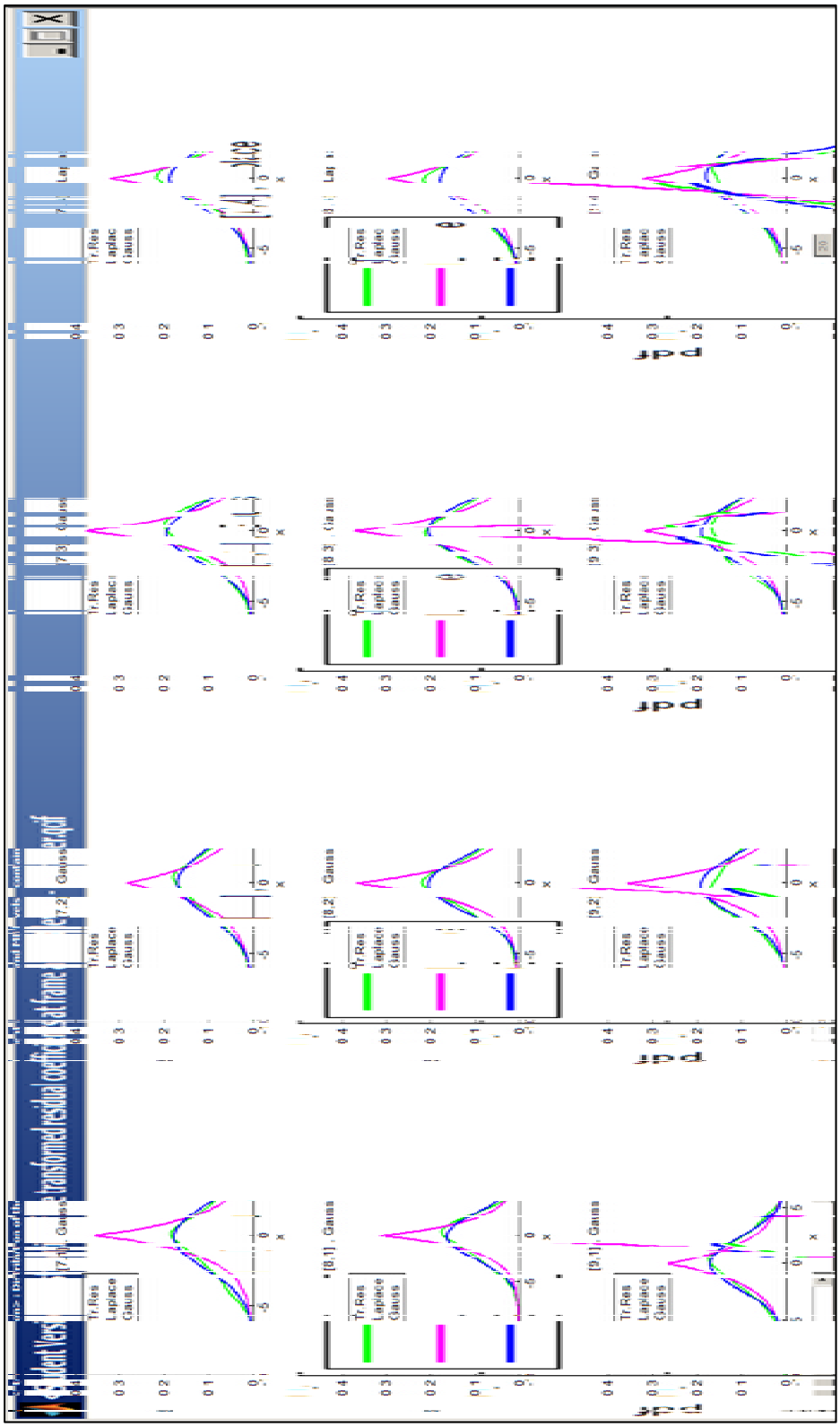


Figure 5.2 Container qcif.yuv, frame 2(P), QP = 20, rendered with mixed (Gauss and Laplace) model

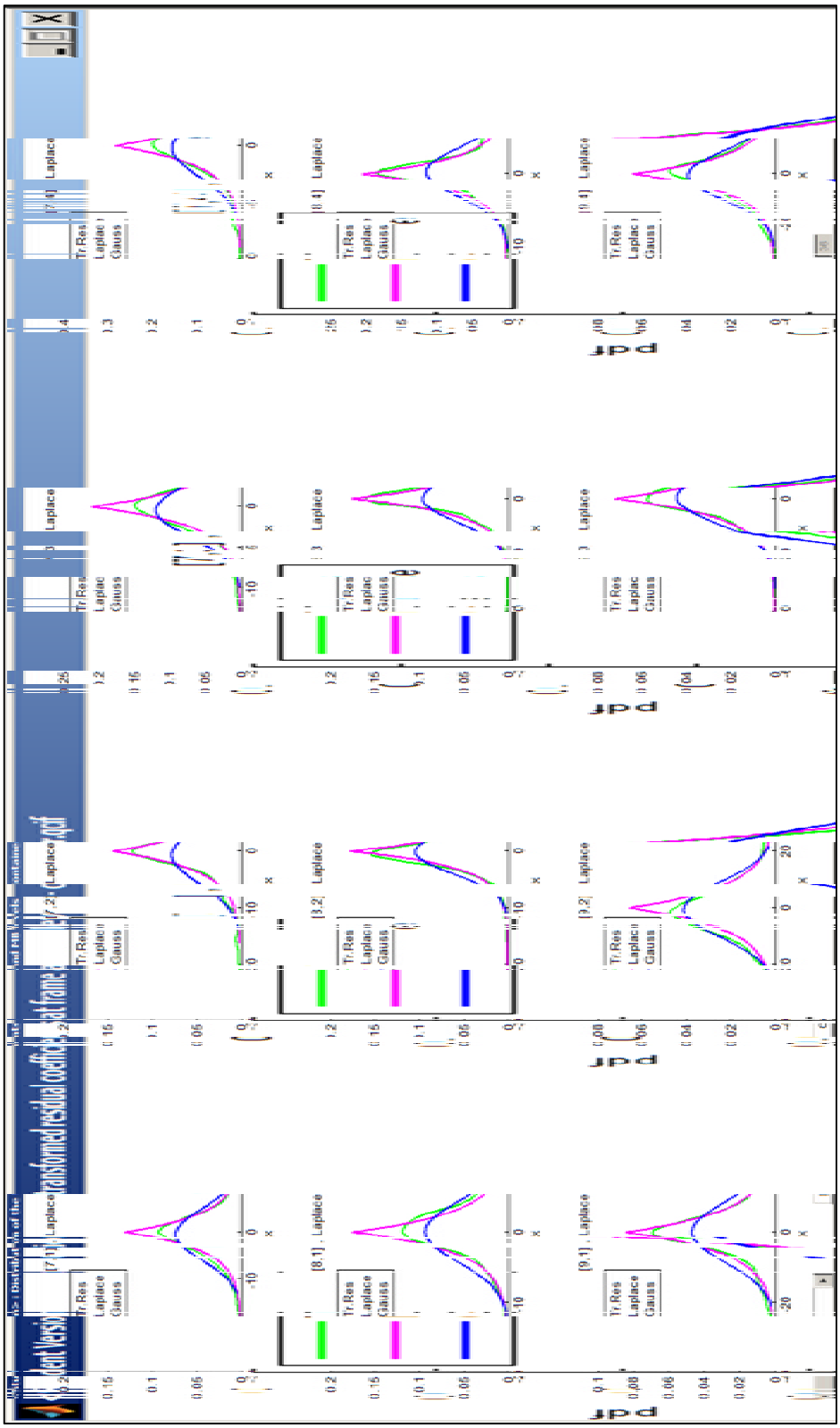


Figure 5.3 Container_qcif.yuv frame 2(P) QP = 36, rendered with mixed (Gauss and Laplace) model

If it were to consider the μ parameter in the *D* and *H* equations, the closed form of *D* and *H* would be difficult to achieve. This is why (Li and others, 2009) preferred to assume a $\mu = 0$ for the expressions of *D* and *H* and make several corrections on the fly in order to ensure the encoding process does keep on the right track.

If the estimated Lagrange multiplier is below the value determined at high rate or if exceeds the upper threshold then it is limited to 0.9 λ_{HR} and respectively up to five times the value of λ_{HR}.

5.3 H.264 rate and distortion estimation at macroblock level

In this set of simulations, the video clips were encoded with λ_{HR} and statistics were gathered to estimate the various models' parameters from residual information and measure how well we could estimate the MB level rate and distortion values. It is important to note that in these simulations, unlike those of the next sub-section, we did not interfere with the regular encoding process by applying a different Lagrange multiplier.

The tables 5.1, 5.2, and A V-1 through A V-10 in the ANNEX V show that the distortion's relative error at the macroblock level, using the real value and the one calculated with each model, has small mean and standard deviation values, under 10%, and decreases with QP for slow (container_qcif, container_cif) and medium-paced sequences (coastguard_qcif, silent_qcif). Nevertheless, for fast sequences, the mean and standard deviation values increase significantly (ice_qcif, ice_cif, and foreman_qcif) for the whole QP domain. This partly explains why the R-D gains are more important for slow sequences, like container. It can also observed that, from the standpoint of relative error of distortion, the mixed Laplace-Gauss model is just a little bit better than Laplace's, but only for slow sequences. On the contrary, the rate relative error with respect to the real rate has huge values increasing with QP. Therefore, accuracy problems are expected when computing the Lagrange multiplier. In summary, for both Laplace and Laplace-Gauss models, the distortion's relative error decreases with an increase of the QP.

Table 5.1 Consolidated statistics of MB models applied to sequence container_cif

Container_cif, Laplace	QP=28	QP=32	QP=36	QP=40
Distortion - rel. err. (%) (mean)	7.36	4.08	2.76	1.79
Distortion - rel. err. (%) (std dev)	9.35	7.34	5.63	3.29
Rate - rel. err. (%) (mean)	32.17	24.65	20.08	13.44
Rate - rel. err. (%) (std dev)	51.40	52.16	50.53	44.84
Container_cif, Laplace-Gauss	QP=28	QP=32	QP=36	QP=40
Distortion - rel. err. (%) (mean)	5.97	3.21	2.29	1.09
Distortion - rel. err. (%) (std dev)	8.08	6.41	5.03	3.09
Rate - rel. err. (%) (mean)	32.05	24.86	20.05	13.51
Rate - rel. err. (%) (std dev)	51.58	52.17	50.51	44.92

Table 5.2 Consolidated statistics of MB models applied to sequence container_qcif

Container_qcif, Laplace	QP=28	QP=32	QP=36	QP=40
Distortion - rel. err. (%) (mean)	8.71	4.99	2.96	1.82
Distortion - rel. err. (%) (std dev)	10.18	7.58	5.35	3.24
Rate - rel. err. (%) (mean)	38.36	29.74	19.62	9.49
Rate - rel. err. (%) (std dev)	52.81	54.34	50.51	39.56
Container_qcif, Laplace-Gauss	QP=28	QP=32	QP=36	QP=40
Distortion - rel. err. (%) (mean)	5.91	3.41	1.71	0.77
Distortion - rel. err. (%) (std dev)	8.02	5.77	3.71	2.64
Rate - rel. err. (%) (mean)	37.22	29.44	19.63	10.43
Rate - rel. err. (%) (std dev)	53.08	54.29	50.51	41.03

The tables 5.3 and 5.4 show that at the MB level, for slow sequences, the new approach based on mixed Laplace-Gauss method, either Laplace or Gauss, benefits from the low values of the distortion prediction error. However, we see clearly that the rate prediction error is very high. Therefore, the rate prediction model is not accurate. Even when the encoding is made with SKIP mode activated, which leads to zero relative error of the distortion of the skipped, when there is a better mode than SKIP, its bitrate relative error is too big to be overridden by a smaller relative error of the distortion. Note that in practice λwill only depend on non-SKIP MBs. Indeed, in the case when SKIP is used, the *R* and *D* derivatives

become zero and non-SKIP MBs control the global value of λ_{mcu}. Even so, we expect that once the MCUs group more than a MB, we see improvements in the bitrate relative error too and lead to more accurate λ_{mcu}.

Table 5.3 The new approach applied to sequence container_qcif

New approach based on Laplace-Gauss model aplied to sequence container_qcif.yuv												
frame	MB	I	c	QP	Real Dist	Model Dist	abs. Dist	rel. Dist(%)	Real Rate	Model Rate	abs. Rate	rel. Rate(%)
7	1	1	1	28	346	448.93	102.07	17.96	0	0	0	0
7	1	1	1	32	1955	1795.4	159.6	10.01	0	0	0	0
7	1	1	1	36	2260	1010.18	360.82	16.22	0	0	0	0
7	1	1	1	40	9772	8802.22	969.78	9.92	0	0	0	0
7	2	1	2	28	541	349.21	191.79	36.17	0	0	0	0
7	2	1	2	32	1981	1770.82	210.18	10.61	0	0	0	0
7	2	1	2	36	2250	1654.83	595.17	27.74	0	0	0	0
7	2	1	2	40	4641	4106.53	534.47	11.52	0	0	0	0
7	3	1	3	28	671	466.94	154.06	24.81	0	0	0	0
7	3	1	3	32	2001	1800.46	203.54	10.16	0	0	0	0
7	3	1	3	36	1840	1081.85	757.15	40.05	0	0	0	0
7	3	1	3	40	5134	4505.05	628.95	12.25	0	0	0	0
7	4	1	4	28	608	387.61	220.39	36.25	0	0	0	0
7	4	1	4	32	1778	1588.29	189.71	10.67	0	0	0	0
7	4	1	4	36	1625	1115.8	509.2	31.34	0	0	0	0
7	4	1	4	40	5344	4452.52	891.48	16.68	0	0	0	0
7	5	1	5	28	2887	1668.75	498.25	18.84	0	0.17	0.17	0
7	5	1	5	32	6073	5385.59	687.41	10.16	0	0.58	-0.58	0
7	5	1	5	36	9457	8052.01	404.99	5.23	0	0.07	-0.07	0
7	5	1	5	40	18528	17300.38	1227.62	6.63	0	0.01	0.01	0
7	6	1	6	28	2673	1488.88	2284.12	29.04	37	20	17	45.93
7	6	1	6	32	15114	10029.83	5084.18	33.64	0	4.91	-4.91	0
7	6	1	6	36	31820	22590.27	9229.73	29	0	4	-4	0
7	6	1	6	40	54239	44312.15	9926.85	18.3	0	1.2	1.2	0
7	7	1	7	28	742	646.48	155.52	19.04	0	0	0	0
7	7	1	7	32	1869	1838.74	538.26	28.91	0	0	0	0
7	7	1	7	36	1942	1514.03	327.97	16.89	0	0	0	0
7	7	1	7	40	4378	3597.35	780.65	17.83	0	0	0	0
7	8	1	8	28	1286	1007.93	278.07	17.58	0	0.04	0.04	0
7	8	1	8	32	3840	3298.99	541.01	14.09	0	0.03	-0.03	0
7	8	1	8	36	5401	5005.29	395.71	7.33	0	0	0	0
7	8	1	8	40	8311	7615.07	695.93	8.37	0	0	0	0
7	9	1	9	28	449	258	188	38.08	0	0	0	0
7	9	1	9	32	914	626.89	287.11	31.41	0	0	0	0
7	9	1	9	36	2179	1855.71	323.29	14.84	0	0	0	0
7	9	1	9	40	2671	1986.34	684.66	25.63	0	0	0	0
7	10	1	10	28	566	388.99	177.01	31.27	0	0	0	0
7	10	1	10	32	840	626.18	213.82	25.45	0	0	0	0
7	10	1	10	36	1737	1384.13	352.87	20.32	0	0	0	0
7	10	1	10	40	1865	1600.65	264.35	14.17	0	0	0	0
7	11	1	11	28	1484	1081.05	402.95	34.59	0	0.01	0.01	0
7	11	1	11	32	2130	1588.09	541.91	25.44	0	0	0	0
7	11	1	11	36	3810	3209.46	600.54	15.76	0	0	0	0
7	11	1	11	40	4211	2310.54	1901.45	44.52	0	0	0	0

Table 5.4 The new approach applied to the sequence silent_qcif.yuv

The new approach based on Laplace model applied to the sequence silent_qcif.yuv												
frame	MB	i	c	QP	Real Dist	Model Dist	abs. Dist	rel. Dist(%)	Real Rate	Model Rate	abs. Rate	rel. Rate(%)
1	1	1	1	28	2440	2004.04	435.96	17.87	0	0.55	0.55	0
7	1	1	1	32	3651	3351.02	299.98	12.28	0	0.03	-0.03	0
7	1	1	1	36	7335	6811.73	618.27	8.49	0	0.01	-0.01	0
1	1	1	1	40	12122	11514.64	607.36	4.97	0	0	0	0
1	2	1	2	28	2357	1867.03	489.97	22.72	0	0.15	0.15	0
7	2	1	2	32	3072	2589.72	482.28	15.7	0	0.01	-0.01	0
7	2	1	2	36	5104	4110.13	993.88	19.47	0	0	0	0
1	2	1	2	40	10568	9968.68	617.32	5.84	0	0	0	0
1	3	1	3	28	2298	1474.15	823.85	37.48	0	0.05	0.05	0
7	3	1	3	32	4117	3335.59	1051.41	20.64	0	0.04	-0.04	0
7	3	1	3	36	7294	5906.79	1387.21	19.02	0	0.01	-0.01	0
1	3	1	3	40	17446	15011.13	1934.87	10.54	0	0.01	0.01	0
1	4	1	4	28	2268	1460.06	842.94	36.81	0	0.07	0.07	0
7	4	1	4	32	3104	2448.12	655.58	21.12	0	0.01	-0.01	0
7	4	1	4	36	3913	2889.77	1035.23	26.19	0	0	0	0
1	4	1	4	40	13726	14890.64	1185.47	7.95	0	0	0	0
1	5	1	5	28	3242	3059.45	182.55	5.63	0	2.14	2.14	0
7	5	1	5	32	7065	5871.12	1193.88	16.85	0	0.54	-0.54	0
7	5	1	5	36	10846	9412.03	1433.98	13.22	0	0.08	-0.08	0
1	5	1	5	40	47184	44041.54	3142.46	6.66	0	0.51	0.51	0
1	6	1	6	28	6823	4500.75	2322.25	32.04	0	9.15	9.15	0
7	6	1	6	32	9665	7957.07	1707.93	17.63	0	1.9	-1.9	0
7	6	1	6	36	16514	14361.67	2152.33	13.03	0	0.58	-0.58	0
1	6	1	6	40	26041	24191.83	1849.17	7.1	0	0.07	0.07	0
1	7	1	7	28	2677	2131.37	545.63	20.38	0	0.47	-0.47	0
7	7	1	7	32	4804	4158.81	645.19	13.43	0	0.11	-0.11	0
7	7	1	7	36	11376	10094.07	1281.93	11.11	0	0.11	-0.11	0
1	7	1	7	40	17424	16324.61	1099.39	6.41	0	0.01	0.01	0
1	8	1	8	28	3628	2588.49	469.51	28.66	0	0.02	-0.02	0
7	8	1	8	32	3477	2907.6	569.4	16.38	0	0.01	-0.01	0
7	8	1	8	36	6135	5168.13	966.88	15.67	0	0	0	0	0
1	8	1	8	40	7427	6718.58	708.42	9.54	0	0	0	0
1	9	1	9	28	1580	1356.47	232.53	14.63	0	0.05	-0.05	0
7	9	1	9	32	2621	2330.06	290.94	11.1	0	0	0	0
7	9	1	9	36	6132	5348.31	783.69	12.78	0	0	0	0
1	9	1	9	40	12078	11544.87	533.13	4.41	0	0	0	0
1	10	1	10	28	1864	1153.23	710.77	38.13	0	0.02	-0.02	0
7	10	1	10	32	3384	2747.05	636.95	18.82	0	0.01	-0.01	0
7	10	1	10	36	6794	6084.28	709.72	10.45	0	0.01	-0.01	0
1	10	1	10	40	14158	12913.3	1244.7	8.8	0	0	0	0
1	11	1	11	28	2822	2237.02	584.98	18.57	0	0.55	-0.55	0
7	11	1	11	32	5425	4650.24	774.76	14.28	0	0.18	-0.18	0
7	11	1	11	36	9306	7797.32	1480.68	16.08	0	0.03	-0.03	0
1	11	1	11	40	17327	15517.4	1809.6	10.44	0	0.01	0.01	0

5.4 H.264 RDO using frame level adaptive at the region level

As described in the section 4.5 the new approach proposes the calculation of the Lagrange multiplier at the frame level using the *R* and *D* derivatives per each MB. The Lagrange multipliers λ_{mcu} computed in this manner is then used to encode each MB in the next frame. The *R* and *D* derivatives can be calculated for two models at the MB level: either with the Laplace model (standalone) or by using the combination of Laplace and Gauss models, a decision that is made when the discrimination criterion (2.32) is employed. One considers a MCU comprised of a single MB, but this new approach has the ability to group multiple MBs having individual σ within a designated range.

The new proposed approach (named here MB LM) and our implementation of the state-of-the-art described in (Li and others, 2009) (named here Frame LM) are compared with the standard implementation JM baseline. The graphics PSNR vs. bitrate (Figures 5.5, 5.6 and A IV-1 through A IV-10 in the ANNEX IV) compare these methods against the JM baseline implementation (blue). The new approach is represented with its two versions: one that uses the Laplace model (red) and a version that employs the mixed model Laplace-Gauss (magenta). The Frame LM method curve is depicted in green. The method of Bjontegaard was used to draw the graphs through four points.

We can observe in the Table 5.5 that the new approach gets several great improvements in terms of BDPSNR for all tested sequences using the standalone Laplace model, except for soccer_qcif, soccer_cif, football_cif, and ice_qcif, where losses are small. The gains of 0.93dB and 0.5dB obtained in the case of container_qcif and container_cif respectively are notable. This fact is accompanied by pronounced bit rate reductions of up to 18.72% and 15.78%, for the same sequences. These BDPSNR gains/BDRATE reductions were made possible by disabling the adaptive rounding. With the combined method Laplace-Gauss we also got BDPSNR gains/BDRATE reductions, although smaller than with Laplace standalone approach: a gain in BDPSNR of 0.31dB and a bit rate reduction of 7.16% in the case of

container_qcif, and a gain in BDPSNR of 0.18dB accompanied by a bit rate reduction of 5.64% for container in CIF format respectively.

Our Frame LM implementation of the state-of-the-art performs well as expected, acquiring a solid gain of 0.88dB along with an impressive rate reduction of almost 19% when encoding the sequence container_qcif. Note that a huge gain of 1.79dB was reported in (Li and others, 2009) during their experiments. But we have some differences in our test setup (e.g. no intra in inter). Furthermore, we could not replicate their results even with the regular JM although we used the same version as them (we could never elucidate this mystery). In the case of Frame LM, one can see that its λ at the frame level has the same values at various encoding QPs as the ones of MCU's model and JM baseline. The curves PSNR vs. bitrate and λ vs. QP of both Laplace and Laplace-Gauss models overlap since for the QP range that we experiment with there is a small number of Gauss type MBs that cannot change the tendency established by the Laplace type MBs.

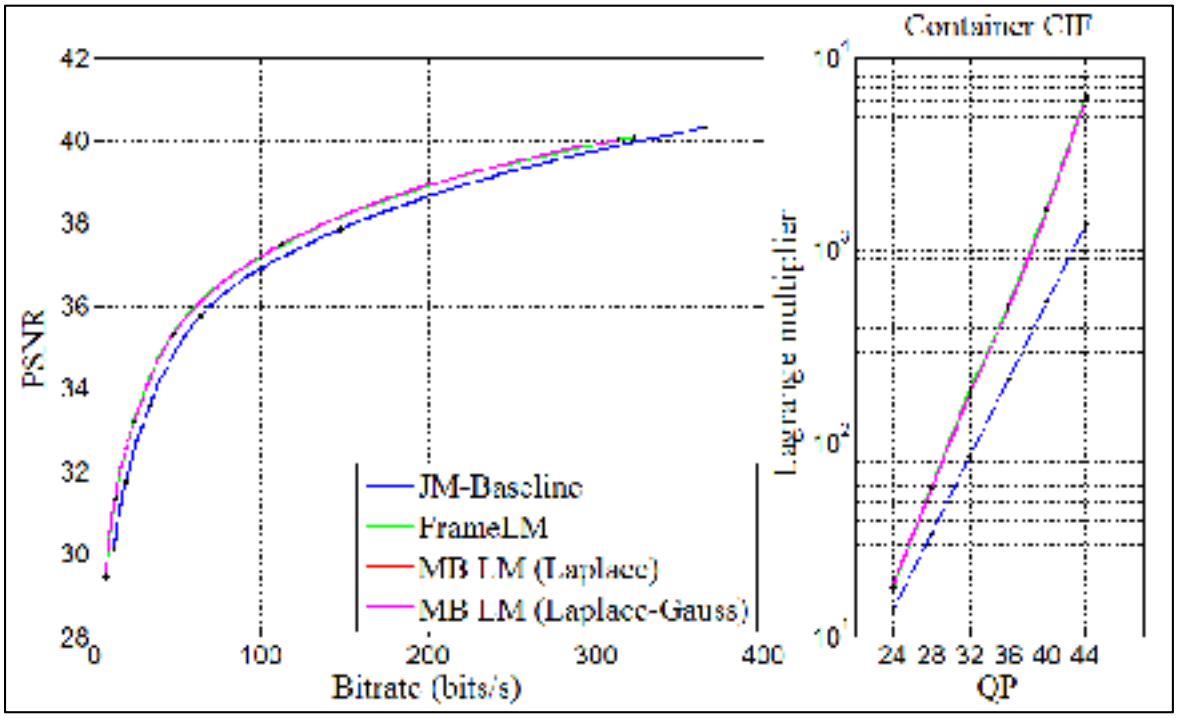


Figure 5.4 PSNR vs. bitrate and λ_{mcu} vs. QP of container_cif.yuv

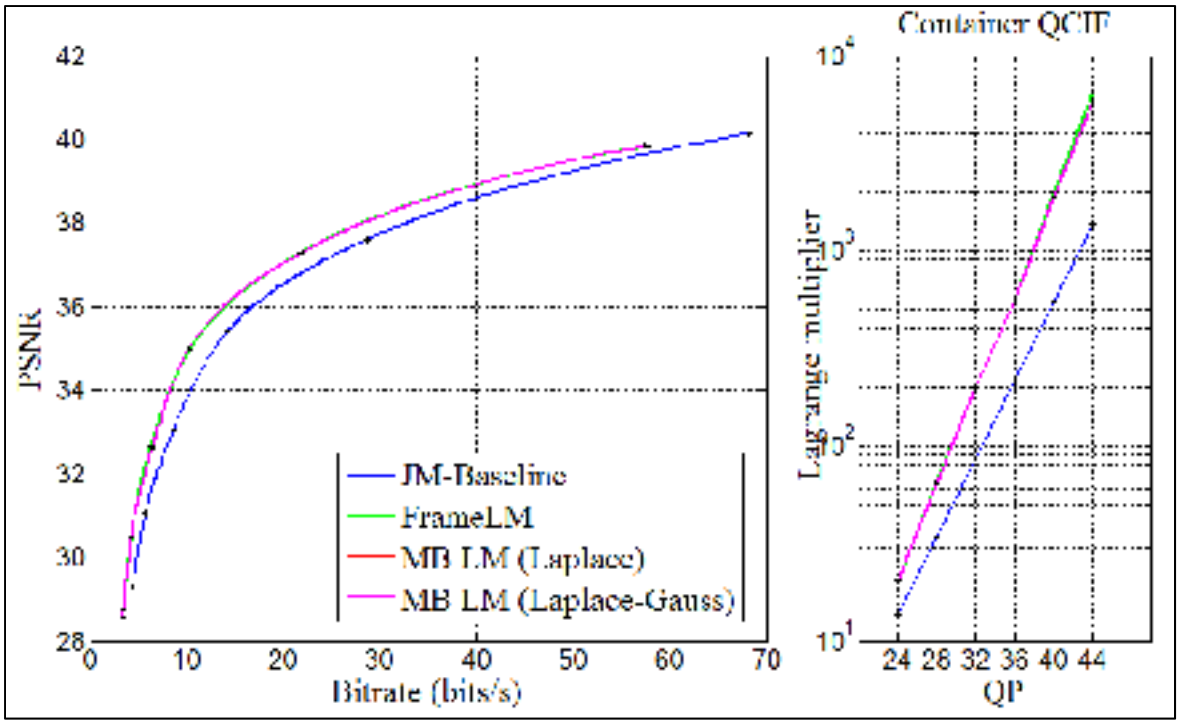


Figure 5.5 PSNR vs. bitrate and λ_{mcu} vs. QP of container_qcif.yuv

The proposed method is comparable from the performance's standpoint with the Frame LM method. Also, they overall perform better than the JM. When MCU approach performs worse than the state-of-the-art method and JM, it is only by a small BDRATE percentage but gains over the JM in the case of container in both QCIF and CIF formats, and silent are noteworthy.

Quasi-static silent and container sequences perform better for Frame LM because the latter computes much higher Lagrange multiplier values which forces more SKIP MBs and reduces dramatically the rate. This makes our rate estimates much less accurate and affects our performance. Thus, we believe that a better rate model would greatly improve the performance of the proposed method. This is a very difficult problem since even (Li and others, 2009) had to rely on several empirical adjustments to obtain a somewhat reliable model (introduction of the S and r parameters and an exponential compensation factor).

Note that the computational complexity of the proposed method is very small compared to the video encoding process since it requires computing simple variance statistics and evaluating a few simple equations per frame.

Table 5.5 Comparison between coding with Laplace at the frame level (FrameLM) and the new approach with respect to the standard implementation of JM

Video	FrameLM BDPSNR	FrameLM BDRate	MB LM (Laplace) BDPSNR	MB LM (Laplace) BDRate	MB LM (Laplace-Gauss) BDPSNR	MB LM (Laplace-Gauss) BDRate
Container_qcif	0.88	18.98	0.93	18.72	0.31	7.16
Foreman_qcif	0.08	1.85	0.08	1.85	0.02	0.49
Silent_qcif	0.38	9.37	0.30	7.64	0.07	1.87
Soccer_qcif	0.11	2.40	0.02	0.40	0.04	0.79
Bus_qcif	0.00	0.12	0.01	0.36	0.00	0.13
Ice_qcif	0.09	1.46	0.03	0.53	0.02	0.37
Coastguard_qcif	0.07	-2.41	0.05	-1.67	0.04	-1.33
Football_qcif	-0.12	2.59	0.01	-0.18	0.02	-0.36
Mobile_cif	0.01	-0.10	0.01	-0.19	-0.02	0.53
Soccer_cif	-0.15	3.96	-0.04	1.25	-0.01	0.68
Container_cif	0.53	-76.41	0.50	-75.78	0.18	-5.64
Football_cif	-0.28	6.04	-0.05	1.06	0.01	-0.30
Coastguard_cif	0.06	-2.02	0.06	-1.96	0.01	0.02
Foreman_cif	-0.01	0.70	0.05	-1.12	-0.02	0.61
Tempete_cif	0.04	-1.11	0.01	-0.03	-0.01	0.30
Ice_cif	-0.24	4.36	-0.05	0.91	0.01	0.01
average	0.06	-7.94	0.12	-2.38	0.04	-0.97

CONCLUSION

Summary of the work

We have proposed a new method for frame level Lagrange multiplier computation based on MB level rate and distortion models.

Although the distortion model nearly attained the values of the real distortion with most video sequences which were tested and especially at higher QP values, the rate, as a byproduct of the motion estimation and mode decision process, did not meet the expectations.

We conjecture that this is because we use a model at the MB level only, that amounts for an insufficient number (256) of samples to correctly predict the distribution of transformed residuals and we try to substitute the transform, quantization, and entropy encoding of each MB coefficient with a global method at the MB level that integrates the distribution of the transform residuals over the available quantization intervals (stated by the uniform scalar quantizer) but does not cover the further stages of run-length and tree/arithmetic coding.

We have partially succeeded to solve the problem of using a single distribution for the whole frame by proposing a new approach (based on MCU) that groups multiple MBs under the same umbrella, as an intermediate case between a processing at the MB level only and frame level processing. The tests were performed using the MB as the MCU and we have showed that such division worked very well (despite the insufficient number of samples) especially for slow paced sequences like container where there are impressive bit rate reductions up to almost 19%.

Contributions

Firstly, we have thoroughly demonstrated the analytical expressions of R, D, and λ for various models: Laplace, Gauss, GGD, etc. This is, to our knowledge, the first work containing detailed derivation of these models. Then, we have proposed a novel MCU-based framework where rate and distortion models for each MCU are used to compute the Lagrange multiplier. The Laplace or Laplace-Gauss based distortion expressions proved to be accurate but only for slow sequences and especially for high QP values. We have implemented the proposed method using the MB as the unit. We have tested the BDRATE/BDPSNR performance of the proposed MCU-based approach achieved by compressing several CIF/QCIF sequences and we observed that it was overall better than the JM (up to 18.72% BDRATE reduction) and on average slightly better than the state-of-the-art algorithm.

Future work

Further research should investigate the performance of the proposed method with larger MCUs and improve the distortion and rate models at such larger MCU level. A more accurate rate model that would extend the current entropy model with the actual encoding process (run length and tree/arithmetic encoding) would be required to improve the performance of the proposed method. The Gauss distribution was assumed to have a rate model similar to the one associated with the Laplace distribution, but this needs to be validated and modified if this is not the case.

ANNEX I

LAPLACE DISTRIBUTION-BASED DISTORTION, RATE, AND LAGRANGE MULTIPLIER

The following shows the formulae deduction for distortion, rate and Lagrange multiplier respectively in the case of a zero-mean Laplace-type signal ($\mu \neq 0$, Λ).

1) The calculation of entropy H

$$P_0 = \int_{-(Q-\gamma Q)}^{Q-\gamma Q} pdf(x)dx = 2\int_0^{Q-\gamma Q} \frac{\Lambda}{2} e^{-\Lambda x} dx \tag{AI.1}$$

$$= \int_0^{Q-\gamma Q} -e^{-\Lambda x} d(-\Lambda x) = \left(-e^{-\Lambda x}\right)\Big|_0^{Q-\gamma Q} = 1 - e^{-(1-\gamma)\Lambda Q}$$

$$P_n = \int_{nQ-\gamma Q}^{(n+1)Q-\gamma Q} pdf(x)dx = \int_{nQ-\gamma Q}^{(n+1)Q-\gamma Q} \frac{\Lambda}{2} e^{-\Lambda |x|} dx = \int_{nQ-\gamma Q}^{(n+1)Q-\gamma Q} \frac{\Lambda}{2} e^{-\Lambda x} dx \tag{AI.2}$$

$$= \int_{nQ-\gamma Q}^{(n+1)Q-\gamma Q} \frac{-1}{2} e^{-\Lambda x} d(-\Lambda x) = \left(\frac{-1}{2} e^{-\Lambda x}\right)\Big|_{nQ-\gamma Q}^{(n+1)Q-\gamma Q} = \frac{1}{2} e^{-n\Lambda Q} e^{\gamma\Lambda Q}\left(1 - e^{-\Lambda Q}\right)$$

$$-2\sum_{n=1}^{\infty} P_n \log_2 P_n = \tag{AI.3}$$

$$= -2\sum_{n=1}^{\infty} \frac{1}{2} e^{-n\Lambda Q} e^{\gamma\Lambda Q}\left(1 - e^{-\Lambda Q}\right) \log_2 \left[\frac{1}{2} e^{-n\Lambda Q} e^{\gamma\Lambda Q}\left(1 - e^{-\Lambda Q}\right)\right] =$$

$$= \frac{-1}{\ln 2} e^{\gamma\Lambda Q}\left(1 - e^{-\Lambda Q}\right) \sum_{n=1}^{\infty} e^{-n\Lambda Q} \ln\left[\frac{1}{2} e^{-n\Lambda Q} e^{\gamma\Lambda Q}\left(1 - e^{-\Lambda Q}\right)\right] =$$

$$= \frac{-e^{\gamma\Lambda Q}(1 - e^{-\Lambda Q})}{\ln 2} \left\{ \sum_{n=1}^{\infty} e^{-n\Lambda Q} \ln \frac{e^{\gamma\Lambda Q}(1 - e^{-\Lambda Q})}{2} + \sum_{n=1}^{\infty} e^{-n\Lambda Q} \ln e^{-n\Lambda Q} \right\}$$

$$= \frac{-e^{\gamma\Lambda Q}\left(1-e^{-\Lambda Q}\right)}{\ln 2}\left(S_1+S_2\right)$$

where

$$S_1 = \sum_{n=1}^{\infty} e^{-n\Lambda Q} \ln \frac{e^{\gamma\Lambda Q}(1-e^{-\Lambda Q})}{2} = \ln\left[\frac{e^{\gamma\Lambda Q}(1-e^{-\Lambda Q})}{2}\right]\sum_{n=1}^{\infty} e^{-n\Lambda Q} = \tag{AI.4}$$

$$= \frac{1}{e^{\Lambda Q}-1} \ln\left[\frac{e^{\gamma\Lambda Q}(1-e^{-\Lambda Q})}{2}\right]$$

and

$$S_2 = \sum_{n=1}^{\infty} e^{-n\Lambda Q} \ln(e^{-n\Lambda Q}) = \sum_{n=1}^{\infty} (-n\Lambda Q) e^{-n\Lambda Q} = \sum_{n=1}^{\infty} \frac{-n\Lambda Q}{e^{n\Lambda Q}} = \tag{AI.5}$$

$$= \lim_{m\to\infty} \sum_{n=1}^{m} \frac{-n\Lambda Q}{e^{n\Lambda Q}} = \lim_{m\to\infty}[\Lambda Q \frac{(1-\frac{1}{e^{m\Lambda Q}})}{e^{\Lambda Q}-1} + \Lambda Q \frac{(\frac{1}{e^{\Lambda Q}} - \frac{1}{e^{m\Lambda Q}})}{e^{\Lambda Q}-1} +$$

$$+ \ldots + \Lambda Q \frac{(\frac{1}{e^{(m-1)\Lambda Q}} - \frac{1}{e^{m\Lambda Q}})}{e^{\Lambda Q}-1}] =$$

$$\lim_{m\to\infty} \frac{\Lambda Q}{e^{\Lambda Q}-1}[(1+\frac{1}{e^{\Lambda Q}}+\frac{1}{e^{2\Lambda Q}}+\ldots+\frac{1}{e^{(m-1)\Lambda Q}}) - \frac{m}{e^{m\Lambda Q}}] =$$

$$= \lim_{m\to\infty} \frac{\Lambda Q}{e^{\Lambda Q}-1}\left[\frac{1-\frac{1}{e^{m\Lambda Q}}}{1-\frac{1}{e^{\Lambda Q}}} - \frac{m}{e^{m\Lambda Q}}\right] = \frac{-\Lambda Q e^{\Lambda Q}}{(e^{\Lambda Q}-1)^2}$$

The formula AI.3 becomes:

$$-2\sum_{n=1}^{\infty} P_n \log_2 P_n = \tag{AI.6}$$

$$= \frac{-e^{\gamma\Lambda Q}(1-e^{-\Lambda Q})}{\ln 2}\left\{\frac{1}{e^{\Lambda Q}-1} \ln \frac{e^{\gamma\Lambda Q}(1-e^{-\Lambda Q})}{2} - \frac{\Lambda Q e^{\Lambda Q}}{(e^{\Lambda Q}-1)^2}\right\} =$$

$$= -\frac{e^{\gamma\Lambda Q}(e^{\Lambda Q}-1)}{(\ln 2)e^{\Lambda Q}}\{\frac{1}{(e^{\Lambda Q}-1)}\ln[\frac{e^{\gamma\Lambda Q}(1-e^{-\Lambda Q})}{2}]-\frac{\Lambda Q e^{\Lambda Q}}{(e^{\Lambda Q}-1)^2}\} =$$

$$= -\frac{1}{\ln 2}e^{-(1-\gamma)\Lambda Q}\left\{\left[\ln(1-e^{-\Lambda Q})+\gamma\Lambda Q-\ln 2\right]-\frac{\Lambda Q}{e^{\Lambda Q}-1}\right\} =$$

$$= \frac{1}{\ln 2}e^{-(1-\gamma)\Lambda Q}\left\{\ln 2-\ln(1-e^{-\Lambda Q})-\gamma\Lambda Q+\frac{\Lambda Q}{1-e^{-\Lambda Q}}\right\}$$

Finally, the entropy calculated with Laplace pdf is:

$$H = \frac{-1}{\ln 2}\left(1-e^{-(1-\gamma)\Lambda Q}\right)\ln\left(1-e^{-(1-\gamma)\Lambda Q}\right)+ \tag{AI.7}$$
$$+\frac{e^{-(1-\gamma)\Lambda Q}}{\ln 2}\left\{\ln 2-\ln\left(1-e^{-\Lambda Q}\right)-\gamma\Lambda Q+\frac{\Lambda Q}{1-e^{-\Lambda Q}}\right\}$$

2) The distortion calculation

$$D = \int_{-(Q-\gamma Q)}^{Q-\gamma Q} x^2\, pdf(x)dx + 2\sum_{n=1}^{\infty}\int_{nQ-\gamma Q}^{(n+1)Q-\gamma Q}(x-nQ)^2\, pdf(x)dx = D_0 + D_n \tag{AI.8}$$

where

$$D_0 = \int_{-(Q-\gamma Q)}^{Q-\gamma Q} x^2\, pdf(x)dx = \int_{-(Q-\gamma Q)}^{Q-\gamma Q} x^2\frac{\Lambda}{2}e^{-\Lambda|x|}dx = 2\int_0^{Q-\gamma Q} x^2\frac{\Lambda}{2}e^{-\Lambda x}dx \tag{AI.9}$$
$$= \Lambda\int_0^{Q-\gamma Q} x^2 e^{-\Lambda x}dx = \Lambda\left[e^{-\Lambda x}\left(\frac{x^2}{-\Lambda}-\frac{2x}{(-\Lambda)^2}+\frac{2}{(-\Lambda)^3}\right)\right]\Big|_0^{Q-\gamma Q} =$$
$$= -e^{-(1-\gamma)\Lambda x}\left[Q^2(1-\gamma)^2+\frac{2Q(1-\gamma)}{\Lambda}+\frac{2}{\Lambda^2}\right]+\frac{2}{\Lambda^2}$$

and

$$D_n = 2\sum_{n=1}^{\infty}\int_{nQ-\gamma Q}^{(n+1)Q-\gamma Q}(x-nQ)^2\, pdf(x)dx = 2\sum_{n=1}^{\infty} I_n \tag{AI.10}$$

$$I_n = \int_{nQ-\gamma Q}^{(n+1)Q-\gamma Q} (x-nQ)^2 \frac{\Lambda}{2} e^{-\Lambda|x|} dx = \int_{nQ-\gamma Q}^{(n+1)Q-\gamma Q} (x-nQ)^2 \frac{\Lambda}{2} e^{-\Lambda x} dx \qquad \text{(AI.11)}$$

$$= \int_{-\gamma Q}^{Q-\gamma Q} y^2 \frac{\Lambda}{2} e^{-\Lambda(y+nQ)} dy = \frac{\Lambda}{2} e^{-n\Lambda Q} \int_{-\gamma Q}^{Q-\gamma Q} y^2 e^{-\Lambda y} dy =$$

$$= \frac{\Lambda}{2} e^{-n\Lambda Q} \left\{ (e^{-\Lambda y}) \left[\frac{y^2}{-\Lambda} - \frac{2y}{(-\Lambda)^2} + \frac{2}{(-\Lambda)^3} \right] \Big|_{-\gamma Q}^{Q-\gamma Q} \right\} =$$

$$= \frac{-1}{2} e^{-n\Lambda Q} \left\{ (e^{-\Lambda y}) \left[y^2 + \frac{2y}{\Lambda} + \frac{2}{\Lambda^2} \right] \Big|_{-\gamma Q}^{Q-\gamma Q} \right\} =$$

$$\frac{-1}{2} e^{-n\Lambda Q} \left\{ e^{-\Lambda Q(1-\gamma)} \left[Q^2(1-\gamma)^2 + \frac{2Q(1-\gamma)}{\Lambda} + \frac{2}{\Lambda^2} \right] - e^{\gamma\Lambda Q} \left[\gamma^2 Q^2 - \frac{2\gamma Q}{\Lambda} + \frac{2}{\Lambda^2} \right] \right\}$$

$$= ce^{-n\Lambda Q}$$

AI.10 becomes:

$$D_n = 2\sum_{n=1}^{\infty} I_n = 2c\sum_{n=1}^{\infty} e^{-n\Lambda Q} = 2c\frac{e^{-\Lambda Q}}{1-e^{-\Lambda Q}} = \frac{-2c}{1-e^{\Lambda Q}} = \qquad \text{(AI.12)}$$

$$= \frac{-2}{1-e^{\Lambda Q}} \left(\frac{-1}{2}\right) \left\{ \begin{array}{l} e^{-\Lambda Q(1-\gamma)} \left[Q^2(1-\gamma)^2 + \frac{2Q(1-\gamma)}{\Lambda} + \frac{2}{\Lambda^2} \right] - \\ -e^{\gamma\Lambda Q} \left[\gamma^2 Q^2 - \frac{2\gamma Q}{\Lambda} + \frac{2}{\Lambda^2} \right] \end{array} \right\} =$$

$$= \frac{e^{\gamma\Lambda Q}}{1-e^{\Lambda Q}} \left\{ e^{-\Lambda Q} \left[Q^2(1-\gamma)^2 + \frac{2Q(1-\gamma)}{\Lambda} + \frac{2}{\Lambda^2} \right] - \left[\gamma^2 Q^2 - \frac{2\gamma Q}{\Lambda} + \frac{2}{\Lambda^2} \right] \right\}$$

The distortion final expression is:

$$D = D_0 + D_n = -e^{-(1-\gamma)\Lambda x} \left[Q^2(1-\gamma)^2 + \frac{2Q(1-\gamma)}{\Lambda} + \frac{2}{\Lambda^2} \right] + \frac{2}{\Lambda^2} + \qquad \text{(AI.13)}$$

$$+\frac{e^{\gamma\Lambda Q}}{1-e^{\Lambda Q}}\left\{e^{-\Lambda Q}\left[Q^2(1-\gamma)^2+\frac{2Q(1-\gamma)}{\Lambda}+\frac{2}{\Lambda^2}\right]-\left[\gamma^2Q^2-\frac{2\gamma Q}{\Lambda}+\frac{2}{\Lambda^2}\right]\right\}=$$

$$=\frac{\Lambda Q e^{\gamma\Lambda Q}(2+\Lambda Q-2\gamma\Lambda Q)+2-2e^{\Lambda Q}}{\Lambda^2(1-e^{\Lambda Q})}$$

The distortion calculated for the SKIP mode depends on Λ parameter only.

$$D_{SKIP}=\int_{-\infty}^{\infty}x^2.pdf(x).dx=\int_{-\infty}^{\infty}x^2\frac{\Lambda}{2}e^{-\Lambda|x|}dx= \tag{AI.14}$$

$$\int_0^{\infty}x^2\frac{\Lambda}{2}e^{-\Lambda.x}.dx+\int_{-\infty}^{0}x^2\frac{\Lambda}{2}e^{\Lambda.x}.dx==2.\int_0^{\infty}x^2\frac{\Lambda}{2}e^{-\Lambda.x}.dx=\frac{2}{\Lambda^2}$$

3) The Lagrange multiplier calculation

According to the article (Li and others, 2009), a logarithmic relationship between *R* and *H* exists between rate and entropy.

$$R=S.H.e^{-\zeta\Lambda Q}=\frac{Se^{-\xi.\Lambda.Q}}{\ln 2}.\begin{Bmatrix}-\left(1-e^{-(1-\gamma)\Lambda Q}\right)\ln\left(1-e^{-(1-\gamma)\Lambda Q}\right)+\\ e^{-(1-\gamma)\Lambda Q}\left(\ln 2-\ln(1-e^{-\Lambda Q})-\gamma\Lambda Q+\frac{\Lambda Q}{1-e^{-\Lambda Q}}\right)\end{Bmatrix} \tag{AI.15}$$

The derivatives of distortion and rate are:

$$\frac{\partial D}{\partial Q}=\frac{\Lambda^3\left\{e^{\gamma\Lambda Q}\left[2+2\Lambda Q(1-\gamma)+\gamma\Lambda^2Q^2(1-2\gamma)\right]-2e^{\Lambda Q}\right\}(1-e^{\Lambda Q})}{[\Lambda^2(1-e^{\Lambda Q})]^2}+ \tag{AI.16}$$

$$+\frac{e^{\Lambda Q}\left\{e^{\gamma\Lambda Q}\left[2\Lambda Q+\Lambda^2Q^2(1-2\gamma)\right]+2-2e^{\Lambda Q}\right\}}{[\Lambda^2(1-e^{\Lambda Q})]^2}$$

$$\frac{\partial R}{\partial Q} = \frac{Se^{-\xi\Lambda Q}}{\ln 2}\{(-\xi\Lambda)\{-(1-e^{-(1-\gamma)\Lambda Q})\ln(1-e^{-(1-\gamma)\Lambda Q})+ \quad \text{(AI.17)}$$

$$e^{-(1-\gamma)\Lambda Q}[\ln 2 - \ln(1-e^{-\Lambda Q}) - \gamma\Lambda Q + \frac{\Lambda Q}{1-e^{-\Lambda Q}}]\}$$

$$+\{-\Lambda(1-\gamma)e^{-(1-\gamma)\Lambda Q}[1+\ln(1-e^{-(1-\gamma)\Lambda Q})]-$$

$$\Lambda(1-\gamma)e^{-(1-\gamma)\Lambda Q}[\ln 2 - \ln(1-e^{-\Lambda Q}) - \gamma\Lambda Q + \frac{\Lambda Q}{1-e^{-\Lambda Q}}]+$$

$$e^{-(1-\gamma)\Lambda Q}[\frac{-\Lambda e^{-\Lambda Q}}{1-e^{-\Lambda Q}} - \gamma\Lambda + \frac{\Lambda[1-e^{-\Lambda Q} - \Lambda Q e^{-\Lambda Q}]}{(1-e^{-\Lambda Q})^2}]\}\}$$

Finally, we calculate the Lagrange multiplier expression

$$\lambda_{LAP} = -\frac{dD}{dR} = -\frac{\frac{\partial D}{\partial Q}}{\frac{\partial R}{\partial Q}} \quad \text{(AI.18)}$$

ANNEX II

GAUSS DISTRIBUTION-BASED DISTORTION, RATE, AND LAGRANGE MULTIPLIER

In the case of a Gauss-type signal (μ , σ) the distortion, entropy and Lagrange multiplier depend on μ , σ, and Q step of the uniform quantizer.

1) The entropy calculation

$$H = H_n^- + H_0 + H_n^+ = -\sum_{n=1}^{\infty} P_n^- \log_2 P_n^- - P_0 \log_2 P_0 - \sum_{n=1}^{\infty} P_n^+ \log_2 P_n^+ \tag{AII.1}$$

where:

$$P_0 = \int_{-(Q-\gamma Q)}^{Q-\gamma Q} pdf(x)dx = \int_{-(Q-\gamma Q)}^{Q-\gamma Q} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx = \frac{1}{2} erf\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\Big|_{-Q(1-\gamma)}^{Q(1-\gamma)} \tag{AII.2}$$

$$= \frac{1}{2}\left\{erf\left(\frac{Q(1-\gamma)-\mu}{\sigma\sqrt{2}}\right) - erf\left(\frac{-Q(1-\gamma)-\mu}{\sigma\sqrt{2}}\right)\right\}$$

$$= \frac{1}{2}\left\{erf\left(\frac{Q(1-\gamma)-\mu}{\sigma\sqrt{2}}\right) + erf\left(\frac{Q(1-\gamma)+\mu}{\sigma\sqrt{2}}\right)\right\}$$

$$P_n^+ = \int_{nQ-\gamma Q}^{(n+1)Q-\gamma Q} pdf(x)dx = \int_{nQ-\gamma Q}^{(n+1)Q-\gamma Q} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx \tag{AII.3}$$

$$= \frac{1}{2} erf\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\Big|_{nQ-\gamma Q}^{(n+1)Q-\gamma Q} = \frac{1}{2}\left\{erf\left(\frac{(n+1)Q-\gamma Q-\mu}{\sigma\sqrt{2}}\right) - erf\left(\frac{nQ-\gamma Q-\mu}{\sigma\sqrt{2}}\right)\right\}$$

$$P_n^- = \int_{-(n+1)Q+\gamma Q}^{-nQ+\gamma Q} pdf(x)dx = \tag{AII.4}$$

$$= \int_{-(n+1)Q+\gamma Q}^{-nQ+\gamma Q} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx = \frac{1}{2} erf\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\Big|_{-(n+1)Q+\gamma Q}^{-nQ+\gamma Q} =$$

$$= \frac{1}{2}\left\{erf\left(\frac{(n+1)Q-\gamma Q+\mu}{\sigma\sqrt{2}}\right) - erf\left(\frac{nQ-\gamma Q+\mu}{\sigma\sqrt{2}}\right)\right\}$$

2) As in the case of Laplace distribution, a logarithmic relationship between *R* and *H* may be considered, with the same values for the constants S (1.982) and ζ (0.35).

$$R = S.H.e^{-\sqrt{2}\zeta Q / \sigma} \tag{AII.5}$$

The derivative of R with respect to Q becomes:

$$\frac{\partial R}{\partial Q} = S.e^{-\sqrt{2}\zeta Q / \sigma}\left(\frac{\partial H}{\partial Q} - \frac{\sqrt{2}\zeta}{\sigma} H\right) \tag{AII.6}$$

Further one calculates the derivative of the entropy

$$\frac{\partial H}{\partial Q} = \frac{\partial H_n^-}{\partial Q} + \frac{\partial H_0}{\partial Q} + \frac{\partial H_n^+}{\partial Q} \tag{AII.7}$$

$$\frac{\partial H_0}{\partial Q} = \frac{-1}{\ln 2}\left[\frac{\partial P_0}{\partial Q}\ln P_0 + P_0\frac{1}{P_0}\frac{\partial P_0}{\partial Q}\right] = \frac{-1}{\ln 2}\frac{\partial P_0}{\partial Q}\left[1+\ln P_0\right] = \tag{AII.8}$$

$$= \frac{-1}{\ln 2}.\frac{(1-\gamma)}{\sigma\sqrt{2\pi}}\left\{e^{-\left[\frac{Q(1-\gamma)-\mu}{\sigma\sqrt{2}}\right]^2} + e^{-\left[\frac{Q(1-\gamma)+\mu}{\sigma\sqrt{2}}\right]^2}\right\}.$$

$$\left\{1+\ln\left[\frac{1}{2}\left[erf\left(\frac{Q(1-\gamma)-\mu}{\sigma\sqrt{2}}\right) + erf\left(\frac{Q(1-\gamma)+\mu}{\sigma\sqrt{2}}\right)\right]\right]\right\}$$

$$\text{where } \frac{\partial P_0}{\partial Q} = \frac{1-\gamma}{\sigma\sqrt{2\pi}}\left\{ e^{-\left(\frac{Q(1-\gamma)-\mu}{\sigma\sqrt{2}}\right)^2} + e^{-\left(\frac{Q(1-\gamma)+\mu}{\sigma\sqrt{2}}\right)^2} \right\}$$

$$\frac{\partial H_n^+}{\partial Q} = \frac{-1}{\ln 2}\sum_{n=1}^{\infty}\left[\frac{\partial P_n^+}{\partial Q}\ln P_n^+ + P_n^+\frac{1}{P_n^+}\frac{\partial P_n^+}{\partial Q}\right] = \frac{-1}{\ln 2}\sum_{n=1}^{\infty}\frac{\partial P_n^+}{\partial Q}\left[1+\ln P_n^+\right] = \tag{AII.9}$$

$$= \frac{-1}{\ln 2}\sum_{n=1}^{\infty}\frac{1}{2}\left\{ \frac{2(n+1-\gamma)}{\sigma\sqrt{2\pi}} e^{-\left[\frac{(n+1)Q-\gamma Q-\mu}{\sigma\sqrt{2}}\right]^2} - \frac{2(n-\gamma)}{\sigma\sqrt{2\pi}} e^{-\left[\frac{nQ-\gamma Q-\mu}{\sigma\sqrt{2}}\right]^2} \right\}.$$

$$.\left\{1+\ln\left[\frac{1}{2}\left[erf\left(\frac{(n+1)Q-\gamma Q-\mu}{\sigma\sqrt{2}}\right)-erf\left(\frac{nQ-\gamma Q-\mu}{\sigma\sqrt{2}}\right)\right]\right]\right\}$$

$$\text{where } \frac{\partial P_n^+}{\partial Q} =$$

$$\frac{1}{2}\left\{ \frac{2(n+1-\gamma)}{\sigma\sqrt{2\pi}} e^{-\left[\frac{(n+1)Q-\gamma Q-\mu}{\sigma\sqrt{2}}\right]^2} - \frac{2(n-\gamma)}{\sigma\sqrt{2\pi}} e^{-\left[\frac{nQ-\gamma Q-\mu}{\sigma\sqrt{2}}\right]^2} \right\}$$

$$\frac{\partial H_n^-}{\partial Q} = \frac{-1}{\ln 2}\sum_{n=1}^{\infty}\left[\frac{\partial P_n^-}{\partial Q}\ln P_n^- + P_n^-\frac{1}{P_n^-}\frac{\partial P_n^-}{\partial Q}\right] = \frac{-1}{\ln 2}\sum_{n=1}^{\infty}\frac{\partial P_n^-}{\partial Q}\left[1+\ln P_n^-\right] = \tag{AII.10}$$

$$= \frac{-1}{\ln 2}\sum_{n=1}^{\infty}\frac{1}{2}\left\{ \frac{2(n+1-\gamma)}{\sigma\sqrt{2\pi}} e^{-\left[\frac{(n+1)Q-\gamma Q+\mu}{\sigma\sqrt{2}}\right]^2} - \frac{2(n-\gamma)}{\sigma\sqrt{2\pi}} e^{-\left[\frac{nQ-\gamma Q+\mu}{\sigma\sqrt{2}}\right]^2} \right\}.$$

$$\cdot\left\{1+\ln\left[\frac{1}{2}\left[erf\left(\frac{(n+1)Q-\gamma Q+\mu}{\sigma\sqrt{2}}\right)-erf\left(\frac{nQ-\gamma Q+\mu}{\sigma\sqrt{2}}\right)\right]\right]\right\}$$

where $\frac{\partial P_n^-}{\partial Q}=$

$$\frac{1}{2}\left\{\frac{2(n+1-\gamma)}{\sigma\sqrt{2\pi}}e^{-\left[\frac{(n+1)Q-\gamma Q+\mu}{\sigma\sqrt{2}}\right]^2}-\frac{2(n-\gamma)}{\sigma\sqrt{2\pi}}e^{-\left[\frac{nQ-\gamma Q+\mu}{\sigma\sqrt{2}}\right]^2}\right\}$$

3) The distortion calculation

$$D=D_n^-+D_0+D_n^+ \tag{AII.11}$$

where

$$D_0=\int_{-(Q-\gamma Q)}^{Q-\gamma Q}x^2\,pdf(x)dx=\int_{-(Q-\gamma Q)}^{Q-\gamma Q}x^2\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}dx \tag{AII.12}$$

$$=\int_A^B(\sigma y+\mu)^2\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}y^2}dy=\ldots$$

$$y=\frac{x-\mu}{\sigma}\quad A=\frac{-Q(1-\gamma)-\mu}{\sigma},B=\frac{Q(1-\gamma)-\mu}{\sigma}$$

$$\ldots=\int_A^B(\sigma^2y^2+2\sigma\mu y+\mu^2)\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}y^2}dy=I_0^1+I_0^2+I_0^3$$

$$I_0^1=\int_A^B\sigma^2y^2\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}y^2}dy=\sigma^2\int_A^By^2\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}y^2}dy=\sigma^2\left[\Phi(y)-y\phi(y)\right]\Big|_A^B \tag{AII.13}$$

$$I_0^2 = 2\mu\sigma\int_A^B y\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}y^2}\,dy = 2\mu\sigma\left[-\phi(y)\right]\Big|_A^B \tag{AII.14}$$

$$I_0^3 = \mu^2\int_A^B \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}y^2}\,dy = \mu^2\left[\Phi(y)\right]\Big|_A^B \tag{AII.15}$$

$$D_0 = \left\{\sigma^2\left[\Phi(y) - y\phi(y)\right] - 2\mu\sigma\phi(y) + \mu^2\Phi(y)\right\}\Big|_A^B = \tag{AII.16}$$

$$= \left\{(\sigma^2+\mu^2)\Phi(y) - \sigma(\sigma y + 2\mu)\phi(y)\right\}\Big|_A^B =$$

$$= \left\{\frac{\sigma^2+\mu^2}{2}erf\left(\frac{y}{\sqrt{2}}\right) - \sigma(\sigma y+2\mu)\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}y^2}\right\}\Big|_A^B =$$

$$= \frac{\sigma^2+\mu^2}{2}\left\{erf\left(\frac{Q(1-\gamma)-\mu}{\sigma\sqrt{2}}\right) + erf\left(\frac{Q(1-\gamma)+\mu}{\sigma\sqrt{2}}\right)\right\} -$$

$$-\frac{\sigma}{\sqrt{2\pi}}\left\{\left[Q(1-\gamma)+\mu\right]e^{-\left[\frac{Q(1-\gamma)-\mu}{\sigma\sqrt{2}}\right]^2} + \left[Q(1-\gamma)-\mu\right]e^{-\left[\frac{Q(1-\gamma)+\mu}{\sigma\sqrt{2}}\right]^2}\right\}$$

$$= D_{01} + D_{02} + D_{03}$$

$$D_n^+ = \sum_{n=1}^{\infty}\int_{nQ-\gamma Q}^{(n+1)Q-\gamma Q}\left(x-nQ\right)^2\,pdf(x)dx = \sum_{n=1}^{\infty} I_n^+ \tag{AII.17}$$

$$I_n^+ = \int_{nQ-\gamma Q}^{(n+1)Q-\gamma Q} (x-nQ)^2 \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx = \ldots \tag{AII.18}$$

with $y = x - nQ + \gamma Q$

$$\ldots = \int_0^Q (y-\gamma Q)^2 \frac{1}{\sigma\sqrt{2\pi}} e^{-\left[\frac{y+nQ-\gamma Q-\mu}{\sigma\sqrt{2}}\right]^2} dy =$$

$$= \int_0^Q \left[(y+nQ-\gamma Q-\mu)-nQ+\mu\right]^2 \frac{1}{\sigma\sqrt{2\pi}} e^{-\left[\frac{y+nQ-\gamma Q-\mu}{\sigma\sqrt{2}}\right]^2} dy = \ldots$$

with $u = \frac{y+nQ-\gamma Q-\mu}{\sigma}$ $\quad A = \frac{nQ-\gamma Q-\mu}{\sigma}$ $\quad B = \frac{(n+1)Q-\gamma Q-\mu}{\sigma}$

$$\ldots = \sigma^2 \int_A^B \left[u - \frac{nQ-\mu}{\sigma}\right]^2 \frac{1}{\sqrt{2\pi}} e^{\frac{-u^2}{2}} du = \sigma^2 (I_{n1}^+ + I_{n2}^+ + I_{n3}^+)$$

$$I_{n1}^+ = \int_A^B u^2 \frac{1}{\sqrt{2\pi}} e^{\frac{-u^2}{2}} du = \left[\Phi(u) - u\phi(u)\right]\Big|_A^B \tag{AII.19}$$

$$I_{n2}^+ = \int_A^B (-2)u\left[\frac{nQ-\mu}{\sigma}\right]\frac{1}{\sqrt{2\pi}} e^{\frac{-u^2}{2}} du = (-2)\left[\frac{nQ-\mu}{\sigma}\right]\int_A^B u \frac{1}{\sqrt{2\pi}} e^{\frac{-u^2}{2}} du = \tag{AII.20}$$

$$= 2\frac{nQ-\mu}{\sigma}\phi(u)\Big|_A^B$$

$$I_{n3}^{+}=\int_{A}^{B}\left[\frac{nQ-\mu}{\sigma}\right]^{2}\frac{1}{\sqrt{2\pi}}e^{\frac{-u^{2}}{2}}du=\left[\frac{nQ-\mu}{\sigma}\right]^{2}\left[\Phi(u)\Big|_{A}^{B}\right] \tag{AII.21}$$

$$I_{n}^{+}=\sigma^{2}\left\{\Phi(u)-u\phi(u)+2\frac{nQ-\mu}{\sigma}\phi(u)+\left[\frac{nQ-\mu}{\sigma}\right]^{2}\Phi(u)\right\}\Big|_{A}^{B}= \tag{AII.22}$$

$$=\sigma^{2}\left\{\left[1+\left(\frac{nQ-\mu}{\sigma}\right)^{2}\right]\Phi(u)+\left[2\left(\frac{nQ-\mu}{\sigma}\right)-u\right]\phi(u)\right\}\Big|_{A}^{B}=$$

$$=\sigma^{2}\left\{\left[\frac{\sigma^{2}+(nQ-\mu)^{2}}{2\sigma^{2}}\right]erf\left(\frac{y+nQ-\gamma Q-\mu}{\sigma\sqrt{2}}\right)+\right.$$

$$\left.+\frac{2(nQ-\mu)-y-nQ+\gamma Q+\mu}{\sigma\sqrt{2\pi}}e^{-\left[\frac{y+nQ-\gamma Q-\mu}{\sigma\sqrt{2}}\right]^{2}}\right\}\Big|_{0}^{Q}=$$

$$=\frac{\sigma^{2}+(nQ-\mu)^{2}}{2}\left\{erf\left[\frac{(n+1)Q-\gamma Q-\mu}{\sigma\sqrt{2}}\right]-erf\left[\frac{nQ-\gamma Q-\mu}{\sigma\sqrt{2}}\right]\right\}+$$

$$+\frac{\sigma}{\sqrt{2\pi}}\left[nQ-Q(1-\gamma)-\mu\right]e^{-\left[\frac{(n+1)Q-\gamma Q-\mu}{\sigma\sqrt{2}}\right]^{2}}-$$

$$-\frac{\sigma}{\sqrt{2\pi}}\left[nQ+\gamma Q-\mu\right]e^{-\left[\frac{nQ-\gamma Q-\mu}{\sigma\sqrt{2}}\right]^{2}}$$

$$D_{n}^{-}=\sum_{n=1}^{\infty}\int_{-(n+1)Q+\gamma Q}^{-nQ+\gamma Q}\left(x+nQ\right)^{2}\,pdf\left(x\right)dx=\sum_{n=1}^{\infty}I_{n}^{-} \tag{AII.23}$$

$$I_n^- = \int_{-(n+1)Q+\gamma Q}^{-nQ+\gamma Q} (x+nQ)^2 \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx = \tag{AII.24}$$

$$= \int_{-Q}^{0} (y+\gamma Q)^2 \frac{1}{\sigma\sqrt{2\pi}} e^{-\left[\frac{y-nQ+\gamma Q-\mu}{\sigma\sqrt{2}}\right]^2} dy = \ldots$$

with $y = x + nQ - \gamma Q$

$$\ldots = \int_{-Q}^{0} \left[(y-nQ+\gamma Q-\mu)+(nQ+\mu)\right]^2 \frac{1}{\sigma\sqrt{2\pi}} e^{-\left[\frac{y-nQ+\gamma Q-\mu}{\sigma\sqrt{2}}\right]^2} dy = \ldots$$

with $u = \frac{y-nQ+\gamma Q-\mu}{\sigma}$ $\quad A = \frac{-Q-nQ+\gamma Q-\mu}{\sigma}$ $\quad B = \frac{-nQ+\gamma Q-\mu}{\sigma}$

$$= \sigma^2 \int_A^B \left[u + \frac{nQ+\mu}{\sigma}\right]^2 \frac{1}{\sqrt{2\pi}} e^{\frac{-u^2}{2}} du = \sigma^2 (I_{n1}^- + I_{n2}^- + I_{n3}^-)$$

$$I_{n1}^- = \int_A^B u^2 \frac{1}{\sqrt{2\pi}} e^{\frac{-u^2}{2}} du = \left[\Phi(u) - u\phi(u)\right]\Big|_A^B \tag{AII.25}$$

$$I_{n2}^- = \int_A^B 2u \left[\frac{nQ+\mu}{\sigma}\right] \frac{1}{\sqrt{2\pi}} e^{\frac{-u^2}{2}} du = 2.\left[\frac{nQ+\mu}{\sigma}\right] \int_A^B u \frac{1}{\sqrt{2\pi}} e^{\frac{-u^2}{2}} du = \tag{AII.26}$$

$$= 2\frac{nQ+\mu}{\sigma}\left[-\phi(u)\right]\Big|_A^B$$

$$I_{n3}^{-}=\int_{A}^{B}\left[\frac{nQ+\mu}{\sigma}\right]^{2}\frac{1}{\sqrt{2\pi}}e^{\frac{-u^{2}}{2}}du=\left[\frac{nQ+\mu}{\sigma}\right]^{2}\left[\Phi(u)\Big|_{A}^{B}\right] \tag{AII.27}$$

$$I_{n}^{-}=\sigma^{2}\left\{\Phi(u)-u\phi(u)-2\frac{nQ+\mu}{\sigma}\phi(u)+\left[\frac{nQ+\mu}{\sigma}\right]^{2}\Phi(u)\right\}\Big|_{A}^{B}= \tag{AII.28}$$

$$=\sigma^{2}\left\{\left[1+\left(\frac{nQ+\mu}{\sigma}\right)^{2}\right]\Phi(u)-\left[2(\frac{nQ+\mu}{\sigma})+u\right]\phi(u)\right\}\Big|_{A}^{B}=$$

$$\sigma^{2}\left\{\begin{array}{l}\left[\frac{\sigma^{2}+(nQ+\mu)^{2}}{2\sigma^{2}}\right]erf\left(\frac{y-nQ+\gamma Q-\mu}{\sigma\sqrt{2}}\right)-\\ -\frac{y+nQ+\gamma Q+\mu}{\sigma\sqrt{2\pi}}e^{-\left[\frac{y-nQ+\gamma Q-\mu}{\sigma\sqrt{2}}\right]^{2}}\end{array}\right\}\Big|_{-Q}^{0}=$$

$$=\frac{\sigma^{2}+(nQ+\mu)^{2}}{2}\left\{erf\left[\frac{(n+1)Q-\gamma Q+\mu}{\sigma\sqrt{2}}\right]-erf\left[\frac{nQ-\gamma Q+\mu}{\sigma\sqrt{2}}\right]\right\}+$$

$$+\frac{\sigma}{\sqrt{2\pi}}\left[nQ-Q(1-\gamma)+\mu\right]e^{-\left[\frac{(n+1)Q-\gamma Q+\mu}{\sigma\sqrt{2}}\right]^{2}}$$

$$-\frac{\sigma}{\sqrt{2\pi}}\left[nQ+\gamma Q+\mu\right]e^{-\left[\frac{nQ-\gamma Q+\mu}{\sigma\sqrt{2}}\right]^{2}}$$

$$D_n = \sum_{n=1}^{\infty}(I_n^- + I_n^+) = \sum_{n=1}^{\infty}(A_1 + A_2 + A_3 + A_4) \tag{AII.29}$$

where:

$$A_1 = \frac{\sigma^2 + (nQ-\mu)^2}{2}\left\{erf\left[\frac{(n+1)Q-\gamma Q-\mu}{\sigma\sqrt{2}}\right] - erf\left[\frac{nQ-\gamma Q-\mu}{\sigma\sqrt{2}}\right]\right\} \tag{AII.30}$$

$$A_2 = \frac{\sigma}{\sqrt{2\pi}}\left\{\begin{array}{l}[nQ-Q(1-\gamma)-\mu]e^{-\left[\frac{(n+1)Q-\gamma Q-\mu}{\sigma\sqrt{2}}\right]^2} - \\ -[nQ+\gamma Q-\mu]e^{-\left[\frac{nQ-\gamma Q-\mu}{\sigma\sqrt{2}}\right]^2}\end{array}\right\}$$

$$A_3 = \frac{\sigma^2 + (nQ+\mu)^2}{2}\left\{erf\left[\frac{(n+1)Q-\gamma Q+\mu}{\sigma\sqrt{2}}\right] - erf\left[\frac{nQ-\gamma Q+\mu}{\sigma\sqrt{2}}\right]\right\}$$

$$A_4 = \frac{\sigma}{\sqrt{2\pi}}\left\{\begin{array}{l}[nQ-Q(1-\gamma)+\mu]e^{-\left[\frac{(n+1)Q-\gamma Q+\mu}{\sigma\sqrt{2}}\right]^2} - \\ -[nQ+\gamma Q+\mu]e^{-\left[\frac{nQ-\gamma Q+\mu}{\sigma\sqrt{2}}\right]^2}\end{array}\right\}$$

4) The calculation of the distortion derivative

$$\frac{\partial D}{\partial Q} = \frac{\partial D_0}{\partial Q} + \frac{\partial D_n}{\partial Q} \tag{AII.31}$$

$$\frac{\partial D_0}{\partial Q} = \frac{\partial D_{01}}{\partial Q} + \frac{\partial D_{02}}{\partial Q} + \frac{\partial D_{03}}{\partial Q} \tag{AII.32}$$

$$\frac{\partial D_n}{\partial Q} = \sum_{n=1}^{\infty}\left(\frac{\partial A_1}{\partial Q} + \frac{\partial A_2}{\partial Q} + \frac{\partial A_3}{\partial Q} + \frac{\partial A_4}{\partial Q}\right) \tag{AII.33}$$

$$\frac{\partial D_{01}}{\partial Q}=\frac{\sigma^2+\mu^2}{2}\left\{\frac{2}{\sqrt{\pi}}e^{-\left[\frac{Q(1-\gamma)-\mu}{\sigma\sqrt{2}}\right]^2}\frac{1-\gamma}{\sigma\sqrt{2}}+\frac{2}{\sqrt{\pi}}e^{-\left[\frac{Q(1-\gamma)+\mu}{\sigma\sqrt{2}}\right]^2}\frac{1-\gamma}{\sigma\sqrt{2}}\right\}= \tag{AII.34}$$

$$=\frac{(1-\gamma)(\sigma^2+\mu^2)}{\sigma\sqrt{2\pi}}\left\{e^{-\left[\frac{Q(1-\gamma)-\mu}{\sigma\sqrt{2}}\right]^2}e^{-\left[\frac{Q(1-\gamma)+\mu}{\sigma\sqrt{2}}\right]^2}\right\}$$

$$\frac{\partial D_{02}}{\partial Q}=-\frac{\sigma}{\sqrt{2\pi}}\{(1-\gamma)e^{-\left[\frac{Q(1-\gamma)-\mu}{\sigma\sqrt{2}}\right]^2}+ \tag{AII.35}$$

$$+\left[Q(1-\gamma)+\mu\right]e^{-\left[\frac{Q(1-\gamma)-\mu}{\sigma\sqrt{2}}\right]^2}(-2)\left[\frac{Q(1-\gamma)-\mu}{\sigma\sqrt{2}}\right]\frac{(1-\gamma)}{\sigma\sqrt{2}}\}=$$

$$=-\frac{(1-\gamma)}{\sigma\sqrt{2\pi}}e^{-\left[\frac{Q(1-\gamma)-\mu}{\sigma\sqrt{2}}\right]^2}\left[Q^2(1-\gamma)^2-\sigma^2-\mu^2\right]$$

$$\frac{\partial D_{03}}{\partial Q}=-\frac{\sigma}{\sqrt{2\pi}}\{(1-\gamma)e^{-\left[\frac{Q(1-\gamma)+\mu}{\sigma\sqrt{2}}\right]^2}+ \tag{AII.36}$$

$$\left[Q(1-\gamma)-\mu\right]e^{-\left[\frac{Q(1-\gamma)+\mu}{\sigma\sqrt{2}}\right]^2}(-2)\left[\frac{Q(1-\gamma)+\mu}{\sigma\sqrt{2}}\right]\frac{(1-\gamma)}{\sigma\sqrt{2}}\}=$$

$$=-\frac{(1-\gamma)}{\sigma\sqrt{2\pi}}e^{-\left[\frac{Q(1-\gamma)+\mu}{\sigma\sqrt{2}}\right]^2}\left[Q^2(1-\gamma)^2-\sigma^2-\mu^2\right]$$

Finally

$$\frac{\partial D_0}{\partial Q}=\frac{Q^2(1-\gamma)^3}{\sigma\sqrt{2\pi}}\left\{e^{-\left[\frac{Q(1-\gamma)-\mu}{\sigma\sqrt{2}}\right]^2}+e^{-\left[\frac{Q(1-\gamma)+\mu}{\sigma\sqrt{2}}\right]^2}\right\} \tag{AII.37}$$

$$\frac{\partial A_1}{\partial Q}=n(nQ-\mu)\left\{erf\left[\frac{(n+1)Q-\gamma Q-\mu}{\sigma\sqrt{2}}\right]-erf\left[\frac{nQ-\gamma Q-\mu}{\sigma\sqrt{2}}\right]\right\}+ \tag{AII.38}$$

$$+\frac{\sigma^2+(nQ-\mu)^2}{2}\left\{\begin{array}{l}\frac{2}{\sqrt{\pi}}\left(\frac{n+1-\gamma}{\sigma\sqrt{2}}\right)e^{-\left[\frac{(n+1)Q-\gamma Q-\mu}{\sigma\sqrt{2}}\right]^2}- \\ -\frac{2}{\sqrt{\pi}}\left(\frac{n-\gamma}{\sigma\sqrt{2}}\right)e^{-\left[\frac{nQ-\gamma Q-\mu}{\sigma\sqrt{2}}\right]^2}\end{array}\right\}=$$

$$=n(nQ-\mu)\left\{erf\left[\frac{(n+1)Q-\gamma Q-\mu}{\sigma\sqrt{2}}\right]-erf\left[\frac{nQ-\gamma Q-\mu}{\sigma\sqrt{2}}\right]\right\}+$$

$$+\frac{\sigma^2+(nQ-\mu)^2}{\sigma\sqrt{2\pi}}\left\{(n+1-\gamma)e^{-\left[\frac{(n+1)Q-\gamma Q-\mu}{\sigma\sqrt{2}}\right]^2}-(n-\gamma)e^{-\left[\frac{nQ-\gamma Q-\mu}{\sigma\sqrt{2}}\right]^2}\right\}$$

$$\frac{\partial A_2}{\partial Q}=\frac{\sigma}{\sqrt{2\pi}}\{[n-(1-\gamma)]e^{-\left[\frac{(n+1)Q-\gamma Q-\mu}{\sigma\sqrt{2}}\right]^2}+ \tag{AII.39}$$

$$+[nQ-Q(1-\gamma)-\mu]e^{-\left[\frac{(n+1)Q-\gamma Q-\mu}{\sigma\sqrt{2}}\right]^2}(\frac{-1}{2})2\left[\frac{(n+1)Q-\gamma Q-\mu}{\sigma}\right]\left[\frac{n+1-\gamma}{\sigma}\right]-$$

$$-(n+\gamma)e^{-\left[\frac{nQ-\gamma Q-\mu}{\sigma\sqrt{2}}\right]^2}-$$

$$(nQ+\gamma Q-\mu)e^{-\left[\frac{nQ-\gamma Q-\mu}{\sigma\sqrt{2}}\right]^2}(\frac{-1}{2})2\left[\frac{nQ-\gamma Q-\mu}{\sigma}\right]\left[\frac{n-\gamma}{\sigma}\right]\}=$$

$$=\frac{\sigma}{\sqrt{2\pi}}\{[(n-1+\gamma)-\frac{n+1-\gamma}{\sigma^2}[(nQ-\mu)^2-Q^2(1-\gamma)^2]]e^{-\left[\frac{(n+1)Q-\gamma Q-\mu}{\sigma\sqrt{2}}\right]^2}+$$

$$+[-(n+\gamma)+\frac{n-\gamma}{\sigma^2}[(nQ-\mu)^2-\gamma^2Q^2]]e^{-\left[\frac{nQ-\gamma Q-\mu}{\sigma\sqrt{2}}\right]^2}\}$$

$$\frac{\partial A_3}{\partial Q}=n(nQ+\mu)\left\{erf\left[\frac{(n+1)Q-\gamma Q+\mu}{\sigma\sqrt{2}}\right]-erf\left[\frac{nQ-\gamma Q+\mu}{\sigma\sqrt{2}}\right]\right\}+ \quad \text{(AII.40)}$$

$$+\frac{\sigma^2+(nQ+\mu)^2}{2}\left\{\begin{array}{l}\frac{2}{\sqrt{\pi}}\left(\frac{n+1-\gamma}{\sigma\sqrt{2}}\right)e^{-\left[\frac{(n+1)Q-\gamma Q+\mu}{\sigma\sqrt{2}}\right]^2}- \\ -\frac{2}{\sqrt{\pi}}\left(\frac{n-\gamma}{\sigma\sqrt{2}}\right)e^{-\left[\frac{nQ-\gamma Q+\mu}{\sigma\sqrt{2}}\right]^2}\end{array}\right\}=$$

$$=n(nQ+\mu)\left\{erf\left[\frac{(n+1)Q-\gamma Q+\mu}{\sigma\sqrt{2}}\right]-erf\left[\frac{nQ-\gamma Q+\mu}{\sigma\sqrt{2}}\right]\right\}+$$

$$+\frac{\sigma^2+(nQ+\mu)^2}{\sigma\sqrt{2\pi}}\left\{(n+1-\gamma)e^{-\left[\frac{(n+1)Q-\gamma Q+\mu}{\sigma\sqrt{2}}\right]^2}-(n-\gamma)e^{-\left[\frac{nQ-\gamma Q+\mu}{\sigma\sqrt{2}}\right]^2}\right\}$$

$$\frac{\partial A_4}{\partial Q}=\frac{\sigma}{\sqrt{2\pi}}\{[n-(1-\gamma)]e^{-\left[\frac{(n+1)Q-\gamma Q+\mu}{\sigma\sqrt{2}}\right]^2}+ \tag{AII.41}$$

$$+[nQ-Q(1-\gamma)+\mu]e^{-\left[\frac{(n+1)Q-\gamma Q+\mu}{\sigma\sqrt{2}}\right]^2}.(\frac{-1}{2})2\left[\frac{(n+1)Q-\gamma Q+\mu}{\sigma}\right]\left[\frac{n+1-\gamma}{\sigma}\right]-$$

$$-(n+\gamma)e^{-\left[\frac{nQ-\gamma Q+\mu}{\sigma\sqrt{2}}\right]^2}-$$

$$-(nQ+\gamma Q+\mu)e^{-\left[\frac{nQ-\gamma Q+\mu}{\sigma\sqrt{2}}\right]^2}(\frac{-1}{2})2\left[\frac{nQ-\gamma Q+\mu}{\sigma}\right]\left[\frac{n-\gamma}{\sigma}\right]\}=$$

$$=\frac{\sigma}{\sqrt{2\pi}}\{[(n-1+\gamma)-\frac{n+1-\gamma}{\sigma^2}[(nQ+\mu)^2-Q^2(1-\gamma)^2]]e^{-\left[\frac{(n+1)Q-\gamma Q+\mu}{\sigma\sqrt{2}}\right]^2}+$$

$$+[-(n+\gamma)+\frac{n-\gamma}{\sigma^2}[(nQ+\mu)^2-\gamma^2Q^2]]e^{-\left[\frac{nQ-\gamma Q+\mu}{\sigma\sqrt{2}}\right]^2}\}$$

D_{SKIP} is calculated when the dead zone extends to the whole domain, so the quantization is not applied to the transformed signal.

$$D_{SKIP,\,Gauss} = \int_{-\infty}^{\infty} x^2\, pdf(x)dx = \int_{-\infty}^{\infty} x^2 . \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left[\frac{x-\mu}{\sigma}\right]^2} .dx = \tag{AII.42}$$

$$= \int_{-\infty}^{\infty} (\mu + \sigma.y)^2 . \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2} .dy = D_{SKIP1} + D_{SKIP2} + D_{SKIP3}$$

$$D_{SKIP1} = \int_{-\infty}^{\infty} \mu^2 . \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2} .dy = \mu^2 . \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}y^2} .dy = \mu^2 \tag{AII.43}$$

$$D_{SKIP2} = \int_{-\infty}^{\infty} 2\mu\sigma.y \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2} .dy = 2\mu\sigma . \int_{-\infty}^{\infty} y \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2} .dy = \tag{AII.44}$$

$$= 2\mu\sigma . \int_{-\infty}^{\infty} y.\phi(y).dy = 2\mu\sigma . \left(-\phi(y)\right)\Big|_{-\infty}^{\infty} = 0$$

$$D_{SKIP3} = \int_{-\infty}^{\infty} \sigma^2 . y^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2} .dy = \sigma^2 . \int_{-\infty}^{\infty} y^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2} .dy = \tag{AII.45}$$

$$= \sigma^2 . \int_{-\infty}^{\infty} y^2 .\phi(y).dy = \sigma^2 . \left[\Phi(y) - y.\phi(y)\right]\Big|_{-\infty}^{\infty} =$$

$$= \sigma^2 . \left[\frac{1}{2} erf\left(\frac{y}{\sqrt{2}}\right)\right]\Big|_{-\infty}^{\infty} = \sigma^2$$

$$D_{SKIP,\,Gauss} = \mu^2 + \sigma^2 \tag{AII.46}$$

5) Finally, we calculate the Lagrange multiplier expression:

$$\lambda_{GAUSS} = -\frac{dD}{dR} = -\frac{\frac{\partial D}{\partial Q}}{\frac{\partial R}{\partial Q}} \tag{AII.47}$$

ANNEX III

GENERALIZED GAUSS DISTRIBUTION-BASED DISTORTION, RATE, AND LAGRANGE MULTIPLIER

With the notation, $t = \left(\frac{\sqrt{2}.Q}{\beta} \right)^{\alpha}$ the expressions of distortion and entropy in (Sun and others, 2013b) become:

$$H_{GG}(t) = -\frac{\ln\left(1-e^{-z.t}\right)}{\ln 2} + e^{-z.t} + \frac{t.e^{-z.t}.\left[z+e^{-t}-z.e^{-t}\right]}{\left(1-e^{-t}\right).\ln 2} \quad \text{(AIII.1)}$$

$$D_{GG}(t) = \beta^2.\left\{1+\frac{e^{-z.t}}{2.\left(1-e^{-t}\right)}\left[t^2.(1-2z)-2t\right]\right\} \quad \text{(AIII.2)}$$

The entropy and distortion derivatives are:

$$t' = \frac{\alpha}{Q}.t \quad \text{(AIII.3)}$$

$$H'_{GG}(t) = \frac{-z.e^{-z.t}.t'}{\left(1-e^{-z.t}\right).\ln 2} - ze^{-z.t}.t' + \quad \text{(AIII.4)}$$

$$+\frac{\left\{\left[t'.e^{-z.t}-t.z.e^{-z.t}.t'\right]\left(z+e^{-t}-z.e^{-t}\right)+t.e^{-z.t}.\left[-e^{-t}.t'+z.e^{-t}.t'\right]\right\}\left(1-e^{-t}\right)}{\left(1-e^{-t}\right)^2.\ln 2} -$$

$$\frac{\left(t.e^{-z.t}\right).\left(z+e^{-t}-z.e^{-t}\right).e^{-t}.t'}{\left(1-e^{-t}\right)^2.\ln 2} = e^{-z.t}.t'\{\frac{-z}{\left(1-e^{-z.t}\right).\ln 2} - z +$$

$$+\frac{\left[z\left(1-t.z\right).\left(1-e^{-t}\right)+\left(1-t\right).e^{-t}\right].\left(1-e^{-t}\right)-t.e^{-t}.\left[z.\left(1-e^{-t}\right)+e^{-t}\right]}{\left(1-e^{-t}\right)^{2}.\ln 2}\}$$

$$D'_{GG}(t)=\frac{\beta^2}{2}.\frac{\left\{-z.e^{-z.t}.t'.\left[t^2.(1-2z)-2t\right]+e^{-z.t}.\left[2.t.(1-2z).t'-2t'\right]\right\}(1-e^{-t})}{\left(1-e^{-t}\right)^{2}}- \tag{AIII.5}$$

$$-\frac{\beta^2}{2}.\frac{e^{-z.t}.\left[t^2.(1-2z)-2t\right].e^{-t}.t'}{\left(1-e^{-t}\right)^{2}}=$$

$$=\frac{\beta^2.e^{-z.t}.t'}{2.\left(1-e^{-t}\right)^{2}}\left\{\left[-t^2.z.(1-2z)+2.t.(1-z)-2\right].(1-e^{-t})-\left[t^2.(1-2z)-2t\right].e^{-t}\right\}$$

The Lagrangian multiplier is calculated using the above derivatives of distortion and entropy.

$$\lambda_{GG}=-\frac{dD_{GG}}{dR_{GG}}=-\frac{\frac{\partial D_{GG}}{\partial Q}}{\frac{\partial R_{GG}}{\partial Q}}=-\frac{\frac{\partial D_{GG}}{\partial t}}{\frac{\partial R_{GG}}{\partial t}} \tag{AIII.6}$$

ANNEX IV

EXPERIMENTAL CURVES PSNR VS. BITRATE

We list below more graphics that show the dependencies of PSNR vs. bitrate and Lagrange multiplier λ_{mcu} vs. QP for several sequences in format CIF/QCIF in the set utilized for experiments. The comparison between sequences in terms of PSNR and bitrate is illustrated in Table 5.5. The results displayed in these figures were obtained with SKIP mode activated and adaptive rounding deactivated.

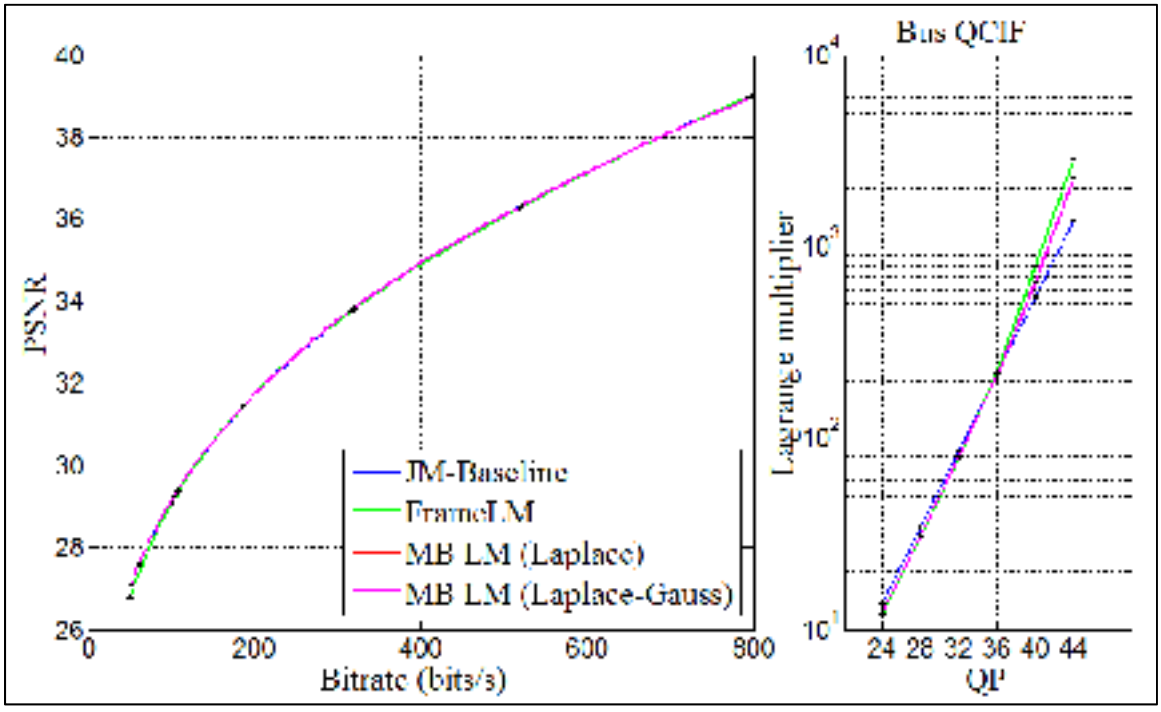


Figure-A IV-1 PSNR vs. bitrate and λ_{mcu} vs. QP of bus_qcif.yuv

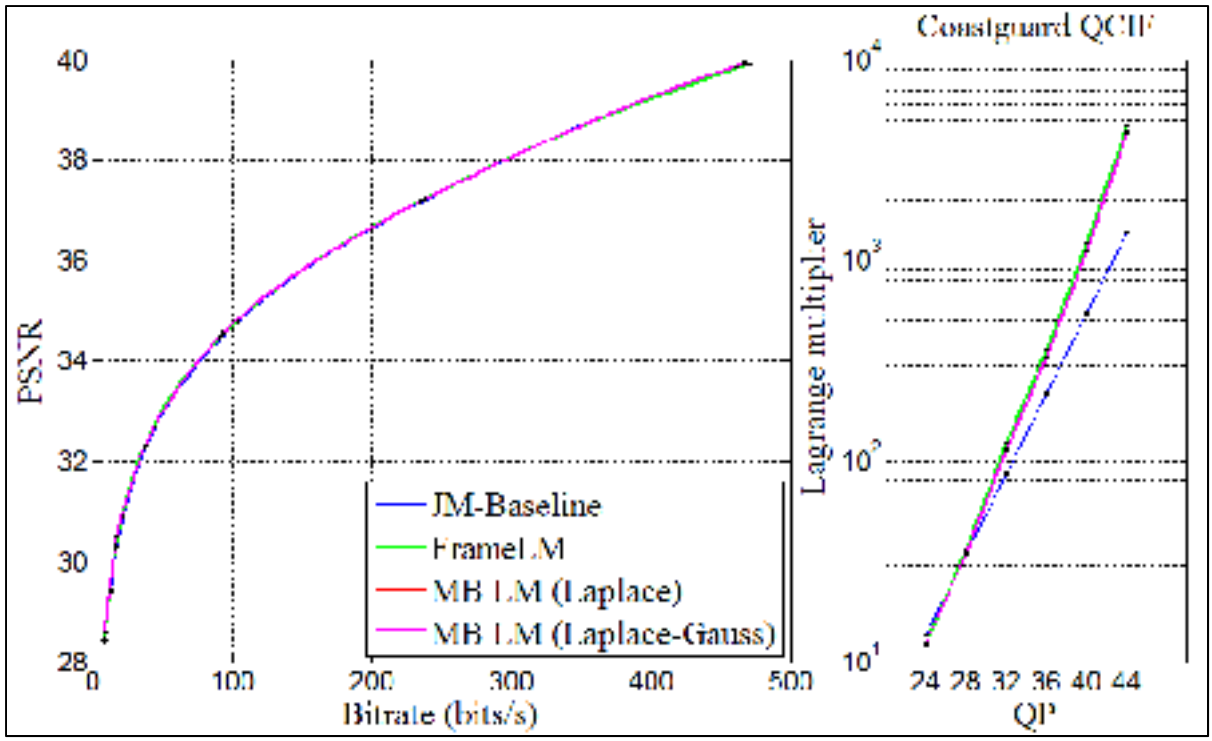


Figure-A IV-2 PSNR vs. bitrate and λ_{mcu} vs. QP of coastguard_qcif.yuv

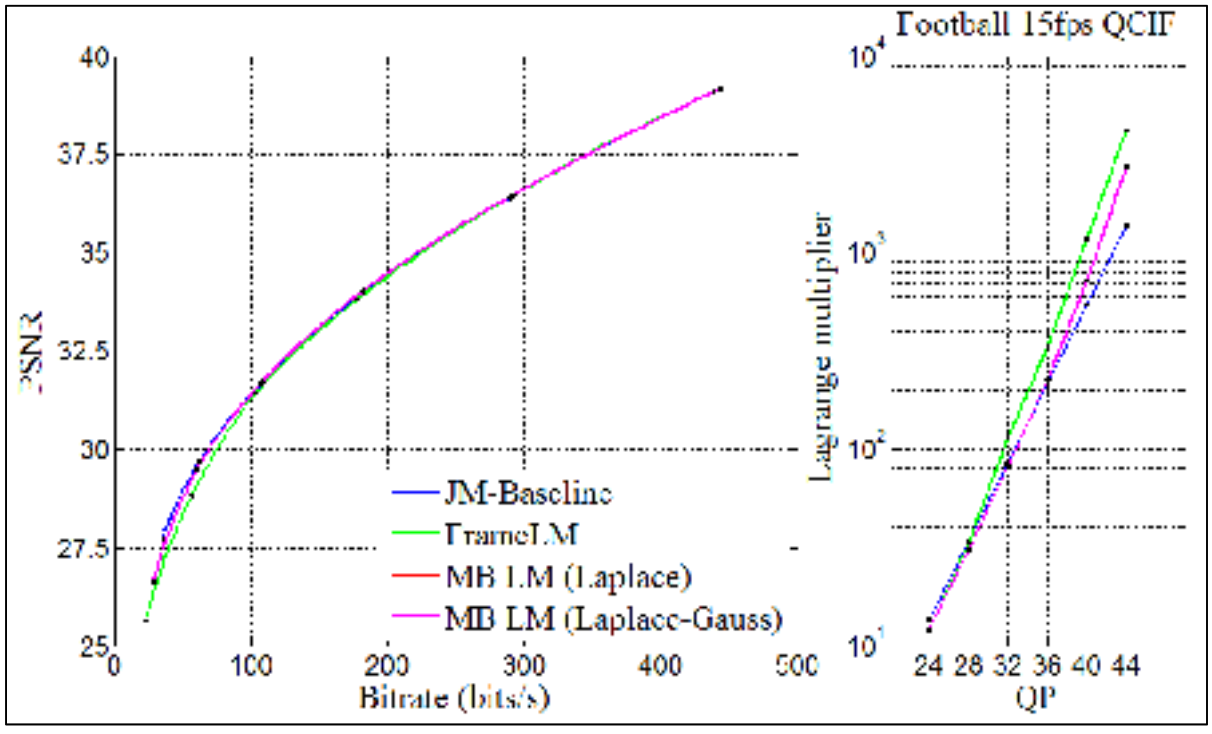


Figure-A IV-3 PSNR vs. bitrate and λ_{mcu} vs. QP of football_qcif.yuv

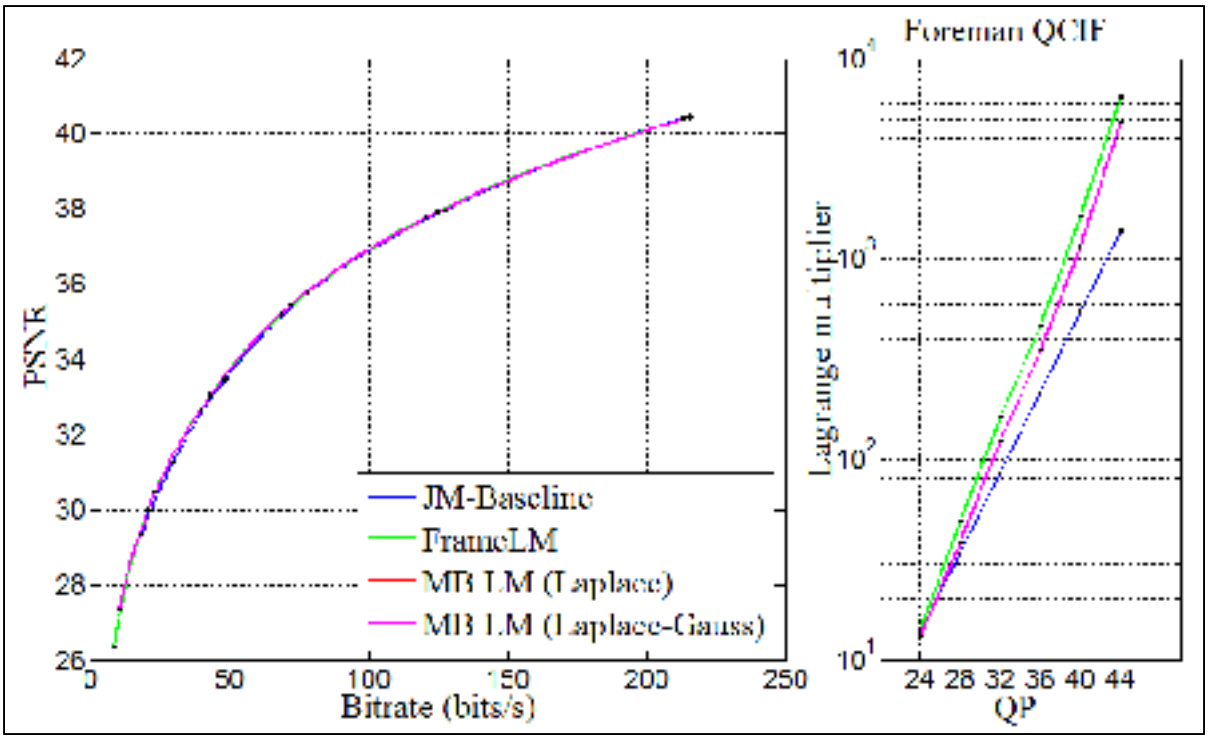


Figure-A IV-4 PSNR vs. bitrate and λ_{mcu} vs. QP of foreman_qcif.yuv

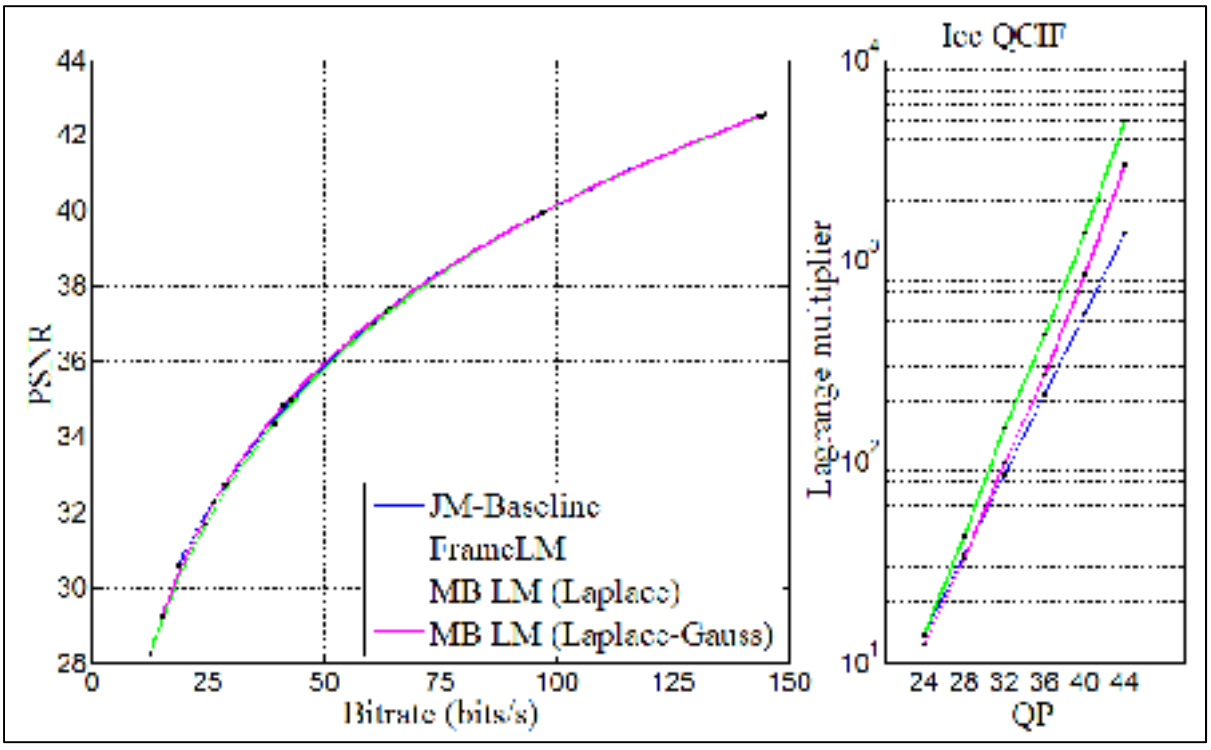


Figure-A IV-5 PSNR vs. bitrate and λ_{mcu} vs. QP of ice_qcif.yuv

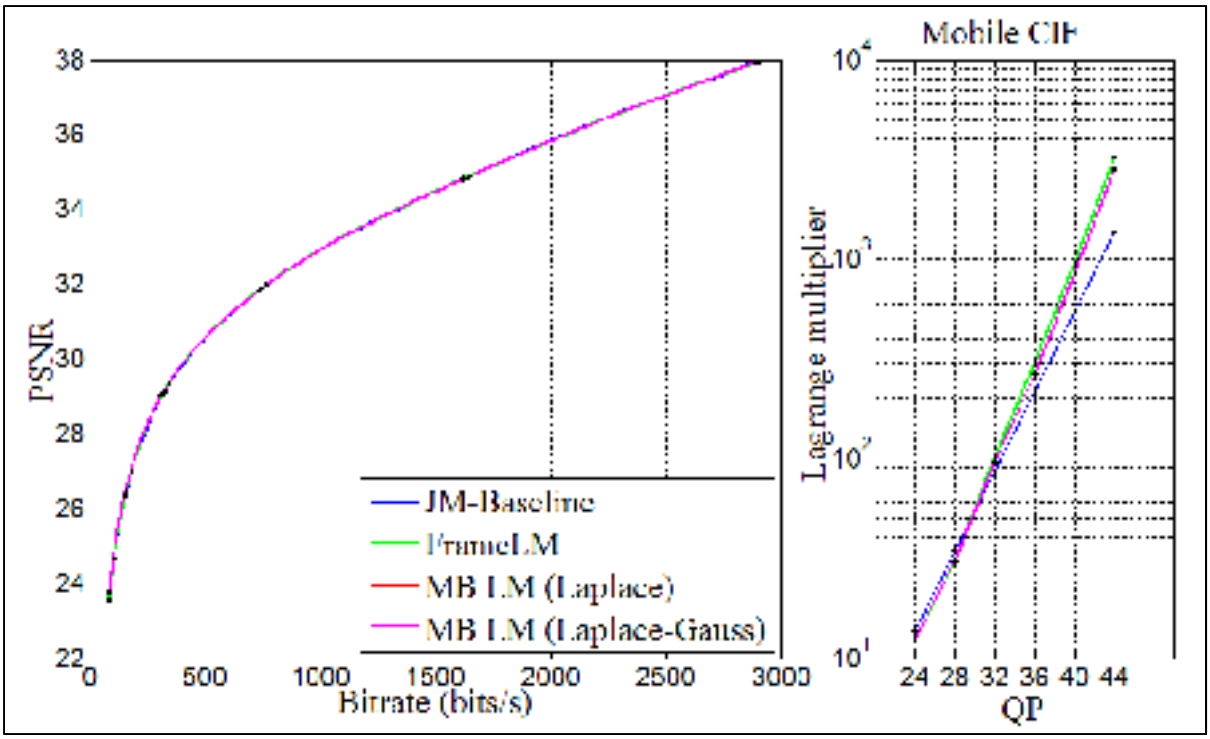


Figure-A IV-6 PSNR vs. bitrate and λ_{mcu} vs. QP of mobile_cif.yuv

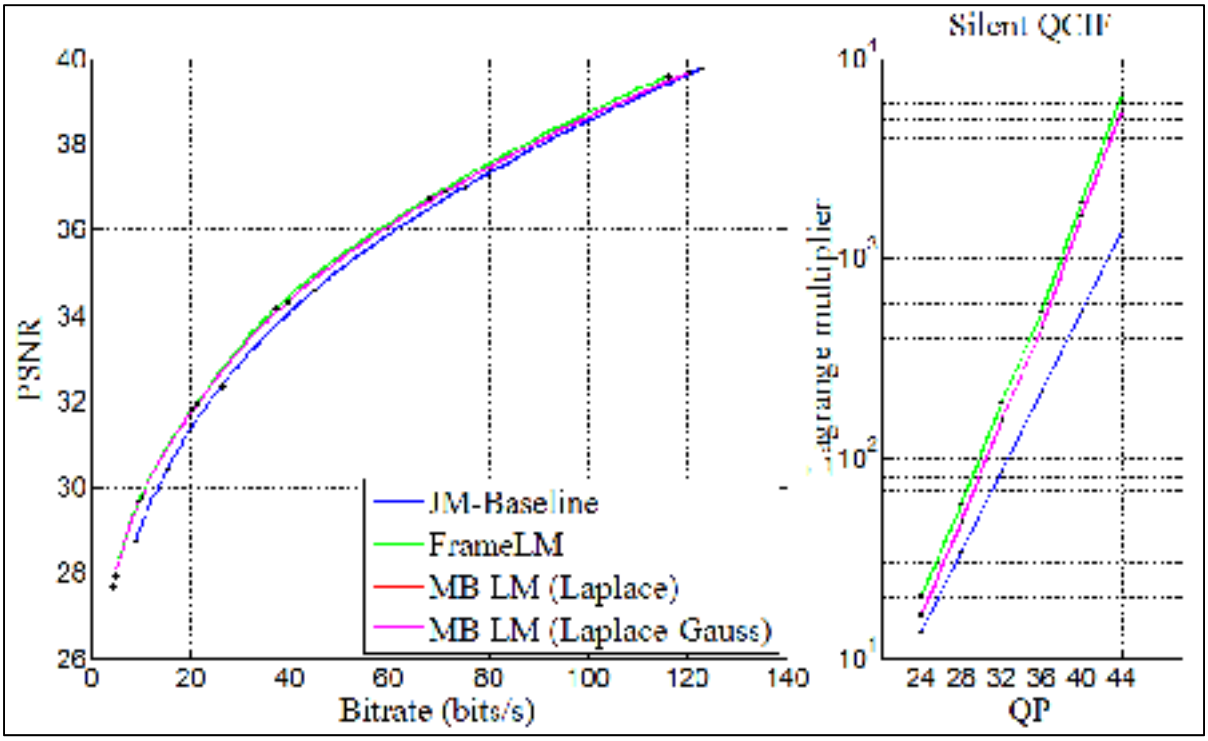


Figure-A IV-7 PSNR vs. bitrate and λ_{mcu} vs. QP of silent_qcif.yuv

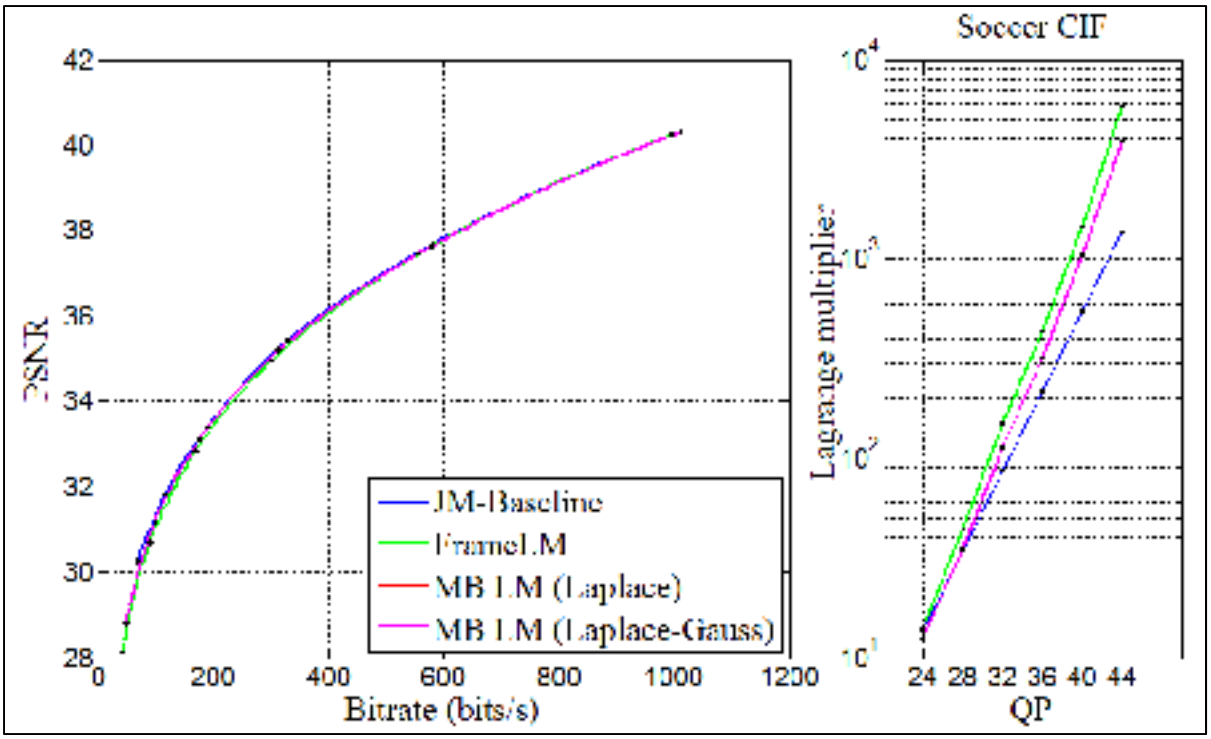


Figure-A IV-8 PSNR vs. bitrate and λ_{mcu} vs. QP of soccer_cif.yuv

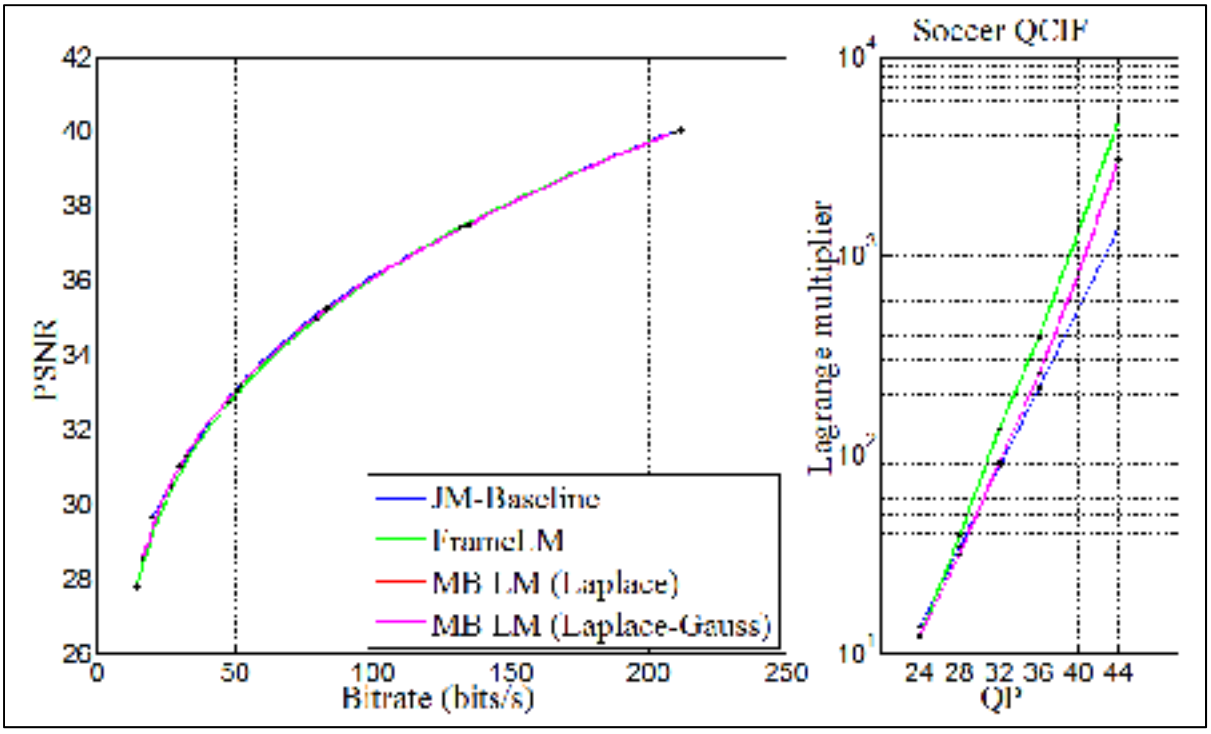


Figure-A IV-9 PSNR vs. bitrate and λ_{mcu} vs. QP of soccer_qcif.yuv

ANNEX V

CONSOLIDATED STATISTICS OF THE NEW APPROACH APPLIED AT THE MACROBLOCK LEVEL

The results displayed in these tables were obtained with SKIP mode activated and adaptive rounding deactivated.

Table-A V- 1 Consolidated statistics of MB models applied to sequence bus_qcif.yuv

bus_qcif, Laplace	QP=28	QP=32	QP=36	QP=40
Distortion - rel. err. (%) (mean)	8.03	9.01	7.54	5.40
Distortion - rel. err. (%) (std dev)	17.99	22.81	22.30	22.02
Rate - rel. err. (%) (mean)	31.96	27.88	22.05	17.96
Rate - rel. err. (%) (std dev)	109.50	92.12	74.47	49.64
bus_qcif, Laplace-Gauss	QP=28	QP=32	QP=36	QP=40
Distortion - rel. err. (%) (mean)	7.82	8.89	7.42	5.37
Distortion - rel. err. (%) (std dev)	17.08	22.80	22.31	22.04
Rate - rel. err. (%) (mean)	31.85	28.02	21.99	17.89
Rate - rel. err. (%) (std dev)	109.35	92.80	74.43	49.35

Table-A V- 2 Consolidated statistics of MB models applied to sequence coastguard_qcif.yuv

Coastguard_qcif, Laplace	QP=28	QP=32	QP=36	QP=40
Distortion - rel. err. (%) (mean)	7.13	4.20	1.76	0.67
Distortion - rel. err. (%) (std dev)	8.56	7.28	4.65	2.28
Rate - rel. err. (%) (mean)	28.04	24.55	15.66	10.55
Rate - rel. err. (%) (std dev)	54.42	50.00	46.47	41.13
Coastguard_qcif, Laplace-Gauss	QP=28	QP=32	QP=36	QP=40
Distortion - rel. err. (%) (mean)	5.90	4.05	1.75	0.68
Distortion - rel. err. (%) (std dev)	8.00	7.19	4.63	2.30
Rate - rel. err. (%) (mean)	27.19	25.27	15.46	10.82
Rate - rel. err. (%) (std dev)	53.14	51.11	46.29	41.51

Table-A V- 3 Consolidated statistics of MB models applied to sequence football_qcif.yuv

football_qcif, Laplace	QP=28	QP=32	QP=36	QP=40
Distortion - rel. err. (%) (mean)	8.59	10.21	8.61	6.77
Distortion - rel. err. (%) (std dev)	16.81	20.31	18.58	17.44
Rate - rel. err. (%) (mean)	29.91	24.46	20.38	18.35
Rate - rel. err. (%) (std dev)	85.60	87.94	68.15	53.83
football_qcif, Laplace-Gauss	QP=28	QP=32	QP=36	QP=40
Distortion - rel. err. (%) (mean)	8.73	10.11	8.57	6.76
Distortion - rel. err. (%) (std dev)	16.83	20.34	18.51	17.44
Rate - rel. err. (%) (mean)	29.91	24.47	20.41	18.35
Rate - rel. err. (%) (std dev)	85.61	87.95	68.17	53.83

Table-A V- 4 Consolidated statistics of MB models applied to sequence foreman_qcif.yuv

foreman_qcif, Laplace	QP=28	QP=32	QP=36	QP=40
Distortion - rel. err. (%) (mean)	11.53	8.14	4.23	1.33
Distortion - rel. err. (%) (std dev)	24.66	22.23	15.54	9.56
Rate - rel. err. (%) (mean)	28.18	25.89	19.91	12.16
Rate - rel. err. (%) (std dev)	61.68	57.75	48.54	43.14
foreman_qcif, Laplace-Gauss	QP=28	QP=32	QP=36	QP=40
Distortion - rel. err. (%) (mean)	11.50	8.12	4.21	1.34
Distortion - rel. err. (%) (std dev)	24.73	22.25	15.56	9.57
Rate - rel. err. (%) (mean)	27.79	25.94	19.95	12.19
Rate - rel. err. (%) (std dev)	60.09	57.78	48.56	43.17

Table-A V- 5 Consolidated statistics of MB models applied to sequence ice_qcif.yuv

ice_qcif, Laplace	QP=28	QP=32	QP=36	QP=40
Distortion - rel. err. (%) (mean)	21.25	21.37	15.91	12.23
Distortion - rel. err. (%) (std dev)	50.17	51.96	42.37	38.19
Rate - rel. err. (%) (mean)	58.06	46.38	31.33	21.21
Rate - rel. err. (%) (std dev)	223.23	211.65	162.00	67.08
ice_qcif, Laplace-Gauss	QP=28	QP=32	QP=36	QP=40
Distortion - rel. err. (%) (mean)	21.25	21.37	15.91	12.23
Distortion - rel. err. (%) (std dev)	50.17	51.96	42.37	38.19
Rate - rel. err. (%) (mean)	58.06	46.38	31.33	21.21
Rate - rel. err. (%) (std dev)	223.23	211.65	162.00	67.08

Table-A V- 6 Consolidated statistics of MB models applied to sequence mobile_cif.yuv

mobile_cif, Laplace	QP=28	QP=32	QP=36	QP=40
Distortion - rel. err. (%) (mean)	8.77	7.11	4.92	2.28
Distortion - rel. err. (%) (std dev)	19.93	17.42	14.02	9.52
Rate - rel. err. (%) (mean)	16.37	21.53	21.81	13.38
Rate - rel. err. (%) (std dev)	57.84	65.55	57.26	44.11
mobile_cif, Laplace Gauss	QP=28	QP=32	QP=36	QP=40
Distortion - rel. err. (%) (mean)	7.97	6.27	4.40	2.27
Distortion - rel. err. (%) (std dev)	19.43	17.01	13.72	9.49
Rate - rel. err. (%) (mean)	16.43	21.54	21.62	13.71
Rate - rel. err. (%) (std dev)	60.57	65.01	57.14	44.52

Table-A V- 7 Consolidated statistics of MB models applied to sequence silent_qcif.yuv

silent_qcif, Laplace	QP=28	QP=32	QP=36	QP=40
Distortion - rel. err. (%) (mean)	7.10	3.10	1.99	0.46
Distortion - rel. err. (%) (std dev)	8.53	5.49	4.00	2.00
Rate - rel. err. (%) (mean)	41.32	27.63	21.05	7.89
Rate - rel. err. (%) (std dev)	52.74	54.16	51.57	36.84
silent_qcif, Laplace Gauss	QP=28	QP=32	QP=36	QP=40
Distortion - rel. err. (%) (mean)	4.94	3.07	1.91	0.45
Distortion - rel. err. (%) (std dev)	6.51	5.43	3.95	1.96
Rate - rel. err. (%) (mean)	42.85	26.58	20.25	7.59
Rate - rel. err. (%) (std dev)	52.02	51.87	51.08	36.24

Table-A V- 8 Consolidated statistics of MB models applied to sequence soccer_cif.yuv

soccer_cif, Laplace	QP=28	QP=32	QP=36	QP=40
Distortion - rel. err. (%) (mean)	10.81	8.12	4.84	2.54
Distortion - rel. err. (%) (std dev)	22.13	20.69	16.72	17.52
Rate - rel. err. (%) (mean)	34.78	27.15	19.03	14.23
Rate - rel. err. (%) (std dev)	94.20	93.21	84.00	45.17
soccer_cif, Laplace-Gauss	QP=28	QP=32	QP=36	QP=40
Distortion - rel. err. (%) (mean)	9.01	7.41	4.76	2.54
Distortion - rel. err. (%) (std dev)	21.83	20.76	16.74	17.54
Rate - rel. err. (%) (mean)	35.16	27.20	19.07	14.25
Rate - rel. err. (%) (std dev)	96.60	93.26	84.03	45.20

Table-A V- 9 Consolidated statistics of MB models applied to sequence soccer_qcif.yuv

soccer_qcif, Laplace	QP=28	QP=32	QP=36	QP=40
Distortion - rel. err. (%) (mean)	11.61	11.77	8.59	5.92
Distortion - rel. err. (%) (std dev)	22.53	28.05	23.93	21.75
Rate - rel. err. (%) (mean)	43.01	32.10	21.59	17.50
Rate - rel. err. (%) (std dev)	146.33	111.98	72.92	49.25
soccer_qcif, Laplace Gauss	QP=28	QP=32	QP=36	QP=40
Distortion - rel. err. (%) (mean)	11.09	11.58	8.67	5.93
Distortion - rel. err. (%) (std dev)	22.66	27.76	23.95	21.77
Rate - rel. err. (%) (mean)	43.12	31.53	21.61	17.52
Rate - rel. err. (%) (std dev)	146.39	110.14	72.96	49.27

Table-A V- 10 Consolidated statistics of MB models applied to sequence ice_cif.yuv

ice_cif, Laplace	QP=28	QP=32	QP=36	QP=40
Distortion - rel. err. (%) (mean)	13.76	13.09	10.56	6.37
Distortion - rel. err. (%) (std dev)	38.28	48.96	43.72	31.71
Rate - rel. err. (%) (mean)	44.05	31.25	23.11	17.32
Rate - rel. err. (%) (std dev)	204.70	121.02	69.07	49.79
ice_cif, Laplace Gauss	QP=28	QP=32	QP=36	QP=40
Distortion - rel. err. (%) (mean)	13.76	13.09	10.56	6.37
Distortion - rel. err. (%) (std dev)	38.28	48.96	43.72	31.71
Rate - rel. err. (%) (mean)	44.05	31.25	23.11	17.32
Rate - rel. err. (%) (std dev)	204.70	121.02	69.07	49.79

ANNEX VI

CONFIGURATION FILE

The following configuration file contains the settings utilized during the experiments with the new Lagrange multiplier

Files

InputFile = "carphone_qcif.yuv", InputHeaderLength = 0, StartFrame = 0,
FramesToBeEncoded = 100, FrameRate = 30.0, SourceWidth = 176,
SourceHeight = 144, SourceResize = 0, OutputWidth = 176, OutputHeight = 144,
TraceFile = "trace_enc.txt", ReconFile = "test_rec.yuv", OutputFile = "test.264",
StatsFile = "stats.dat"

MBLagrangeMultiplier

UseCustomLM = 5 # the approach to calculate lambda
# 1 = standard approach
# 2 = Laplace at MB level
# 3 = Gauss at MB level
# 4 = mixt at MB level
# 5 = integral at MB level
# 6 = GGD at MB level
# 7 = Laplace at frame level

QPOFFSET = 1

FrameLagrangeMultiplier

TSC = 0.3, TPSC = 0.8, TARD = 50.0, TPRD = 10.0, TRGapH = 15.0, TDGapH = 5.0,
TRGap = 7.5, TDGap = 3.0 DisplayLaplaceLambda = 0, WriteLaplaceLambda = 1,
LaplaceLambdaFile = "FrameData.txt"

Encoder Control

ProfileIDC = 66, IntraProfile = 0, LevelIDC = 40, IntraPeriod = 0,
IDRPeriod = 0, AdaptiveIntraPeriod = 1, AdaptiveIDRPeriod = 0,
IntraDelay = 0, EnableIDRGOP = 0, EnableOpenGOP = 0, QPISlice = 28,

QPPSlice = 28, FrameSkip = 0, ChromaQPOffset = 0, DisableSubpelME = 0,
SearchRange = 32, MEDistortionFPel = 0, MEDistortionHPel = 2,
MEDistortionQPel = 2, MDDistortion = 2, SkipDeBlockNonRef = 0,
ChromaMCBuffer = 1, ChromaMEEnable = 0, ChromaMEWeight = 1,
NumberReferenceFrames = 5, PList0References = 0, Log2MaxFNumMinus4 = 0,
Log2MaxPOCLsbMinus4 = -1, GenerateMultiplePPS = 0, SendAUD = 0,
ResendSPS = 2, ResendPPS = 0, MbLineIntraUpdate = 0, RandomIntraMBRefresh = 0

PSlice Mode types

PSliceSkip = 0, PSliceSearch16x16 = 1, PSliceSearch16x8 = 1,
PSliceSearch8x16 = 1, PSliceSearch8x8 = 1, PSliceSearch8x4 = 1,
PSliceSearch4x8 = 1, PSliceSearch4x4 = 1, DisableIntra4x4 = 0,
DisableIntra16x16 = 0, DisableIntraInInter = 1, IntraDisableInterOnly = 0,
Intra4x4ParDisable = 0, Intra4x4DiagDisable = 0, Intra4x4DirDisable = 0,
Intra16x16ParDisable = 0, Intra16x16PlaneDisable = 0, ChromaIntraDisable = 0,
EnableIPCM = 0, DisposableP = 0, DispPQPOffset = 0, PreferDispOrder = 1,
PreferPowerOfTwo = 0, FrmStructBufferLength = 16, ChangeQPFrame = 0,
ChangeQPI = 0, ChangeQPP = 0, ChangeQPB = 0, ChangeQPSI = 0, ChangeQPSP = 0

Output Control, NALs

OutFileMode = 0

Picture based Multi-pass encoding

RDPictureDecision = 0, RDPSliceBTest = 0, RDPictureMaxPassISlice = 1,
RDPictureMaxPassPSlice = 2, RDPictureMaxPassBSlice = 3,
RDPictureFrameQPPSlice = 0, RDPictureFrameQPBSlice = 0,
RDPictureDeblocking = 0, RDPictureDirectMode = 0

Deblocking filter parameters

DFParametersFlag = 0, DFDisableRefISlice = 0, DFAlphaRefISlice = 0,
DFBetaRefISlice = 0, DFDisableNRefISlice = 0, DFAlphaNRefISlice = 0,
DFBetaNRefISlice = 0, DFDisableRefPSlice = 0, DFAlphaRefPSlice = 0,
DFBetaRefPSlice = 0, DFDisableNRefPSlice = 0, DFAlphaNRefPSlice = 0,
DFBetaNRefPSlice = 0

Error Resilience / Slices

SliceMode = 0, SliceArgument = 50, num_slice_groups_minus1 = 0,
slice_group_map_type = 0, slice_group_change_direction_flag = 0,
slice_group_change_rate_minus1 = 85, SliceGroupConfigFileName = "sg0conf.cfg",
UseRedundantPicture = 0, NumRedundantHierarchy = 1, PrimaryGOPLength = 10,
NumRefPrimary = 1

Search Range Restriction / RD Optimization

RestrictSearchRange = 2, RDOptimization = 1, I16RDOpt = 0,
SubMBCodingState = 1, DistortionSSIM = 1, DistortionMS_SSIM = 0,
SSIMOverlapSize = 8, DistortionYUVtoRGB = 0, CtxAdptLagrangeMult = 0,
FastCrIntraDecision = 1, DisableThresholding = 0, SkipIntraInInterSlices = 0,
WeightY = 1, WeightCb = 1, WeightCr = 1

Explicit Lambda Usage

UseExplicitLambdaParams = 0, UpdateLambdaChromaME = 0,
FixedLambdaISlice = 0.1, FixedLambdaPSlice = 0.1,
LambdaWeightISlice = 0.65, LambdaWeightPSlice = 0.68,
LossRateA = 5, LossRateB = 0, LossRateC = 0,
FirstFrameCorrect = 0, NumberOfDecoders = 30, RestrictRefFrames = 0

Additional Stuff

UseConstrainedIntraPred = 0, NumberofLeakyBuckets = 8,
LeakyBucketRateFile = "leakybucketrate.cfg",
LeakyBucketParamFile = "leakybucketparam.cfg",
NumFramesInELayerSubSeq = 0, SparePictureOption = 0,
SparePictureDetectionThr = 6, SparePicturePercentageThr = 92,
PicOrderCntType = 0

#Rate control

RateControlEnable = 0, Bitrate = 45020, InitialQP = 0,
BasicUnit = 0, ChannelType = 0, RCUpdateMode = 0,
RCISliceBitRatio = 1.0, RCBSliceBitRatio0 = 0.5,
RCBSliceBitRatio1 = 0.25, RCBSliceBitRatio2 = 0.25,

RCBSliceBitRatio3 = 0.25, RCBSliceBitRatio4 = 0.25,
RCBoverPRatio = 0.45, RCIoverPRatio = 3.80,
RCMinQPPSlice = 8, RCMaxQPPSlice = 44, RCMinQPISlice = 8,
RCMaxQPISlice = 36

#Fast Mode Decision

EarlySkipEnable = 0
SelectiveIntraEnable = 0
ReportFrameStats = 1
DisplayEncParams = 1
Verbose = 2

#Rounding Offset control

OffsetMatrixPresentFlag = 0, QOffsetMatrixFile = "q_offset.cfg"
AdaptiveRounding = 0, AdaptRoundingFixed = 0, AdaptRndPeriod = 16,
AdaptRndChroma = 1, AdaptRndWFactorIRef = 4, AdaptRndWFactorPRef = 4,
AdaptRndWFactorINRef = 4, AdaptRndWFactorPNRef = 4, AdaptRndCrWFactorIRef = 4,
AdaptRndCrWFactorPRef = 4, AdaptRndCrWFactorINRef= 4,
AdaptRndCrWFactorPNRef= 4

#Fast Motion Estimation Control Parameters

SearchMode = 3, UMHexDSR = 1, UMHexScale = 3, EPZSPattern = 5,
EPZSDualRefinement = 6, EPZSFixedPredictors = 2, EPZSTemporal = 1,
EPZSSpatialMem = 1, EPZSBlockType = 1, EPZSMinThresScale = 0,
EPZSMedThresScale = 1, EPZSMaxThresScale = 2, EPZSSubPelME = 1,
EPZSSubPelMEBiPred = 1, EPZSSubPelThresScale = 2, EPZSSubPelGrid = 1

SEI Parameters

GenerateSEIMessage = 0
SEIMessageText = "H.264/AVC Encoder"
UseMVLimits = 0, SetMVXLimit = 512 , SetMVYLimit= 512
EnableVUISupport = 0

BIBLIOGRAPHY

Altunbasak, Y. and N. Kamaci. 2004. « An analysis of the DCT coefficient distribution with the H.264 video coder ». *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing,* vol. 3, p. iii-177-80.

Atkinson, C. 2012. « YouTube Statistics 2012: A Year Later - More Jawdropping Growth ». < http://www.reelseo.com/youtube-statistics-growth-2012/ (last accessed 07/18/2013) >.

Barker, J. 2011. « Deloitte - Technology, Media & Telecommunications Predictions 2012 ». < http://www.deloitte.com/assets/Dcom-Iceland/Local%20Assets/Documents/TMT%20Prediction%202012.pdf (last accessed 07/18/2013) >.

Bjontegaard, G. 2001. « Calculation of average PSNR differences between RD-curves ». In *13th VCEG Meeting*. (Austin, Texas). < wftp3.itu.int/av-arch/video-site/0104_Aus/VCEG-M33.doc (last accessed 07/18/2013) >.

Bouras, C., and others. 2009. « Video Transmission over TFRC using cross-layer power management ». *SoftCOM 2009*, p. 333 - 337.

Carr, A. 2010. « Blockbuster bankruptcy: a decade of decline ». < http://www.fastcompany.com/1690654/blockbuster-bankruptcy-decade-decline (last accessed 07/18/2013) >.

Chen, Z and K Ngan. 2007. « Recent advances in rate control for video coding ». *Signal Processing: Image Communication,* vol. 22, p. 19-38.

Etherington, D. 2012. « Netflix Hits 30 Million Members ». < http://techcrunch.com/2012/10/25/netflix-hits-30-million-members-after-q3-subscriber-growth-forecasting-error/ (last accessed 07/18/2013) >.

Kovach, S. 2013. « Time Spent Watching Video On Mobile Devices Has Doubled ». < http://www.businessinsider.com/chart-of-the-day-mobile-video-growth-2013-6 (last accessed 07/18/2013) >.

Kundu, D. 2005. « Discriminating Between The Normal and The Laplace Distributions ». *Advances in Ranking and Selection, Multiple Comparisons, and Reliability*, p. 65-79.

Lam, E. Y. and J. W. Goodman. 2000. « A mathematical analysis of the DCT coefficient distributions for images. ». *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society,* vol. 9, p. 1661-6.

Li, X., and others. 2007. « Advanced Lagrange multiplier selection for hybrid video coding ». *IEEE International Conference on Multimedia and Expo*, p. 364-367.

Li, X., and others. 2009. « Laplace Distribution Based Lagrangian Rate Distortion Optimization for Hybrid Video Coding ». *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 19, p. 193-205.

Li, X., N. Oertel, and A. Kaup. 2007. « Adaptive Lagrange multiplier selection for intra-frame video coding ». *ISCAS 2007*, p. 3643-3646.

Puig, P. and M. Stephens. 2007. « Goodness of fit tests for the skew-Laplace distribution ». *Sort: Statistics and Operations Research Transactions,* vol. 31, p. 45-54.

Puig, P. and M. Stephens. 2000. « Tests of Fit for the Laplace Distribution, With Applications ». *Technometrics,* vol. 42, n^{o} 4, p. 417-424.

Richardson, I. 2010. *The H.264 advanced video compression standard*, 2nd. John Wiley & Sons, Ltd.

Stone, B. 2013. « Amazon Said to Plan TV Set-Top Box for Streaming Video ». < http://www.bloomberg.com/news/2013-04-24/amazon-said-to-plan-tv-set-top-box-for-streaming-video.html (last accessed 07/18/2013) >.

Sühring, K. 2013. « H.264/AVC Reference Software Version JM18.3 Joint Video Team. ». < http://iphome.hhi.de/suehring/tml/ >.

Sullivan, G.J. and T. Wiegand. 1998. « Rate-distortion optimization for video compression ». *IEEE Signal Processing Magazine,* vol. 15, p. 74-90.

Sun, J., and others. 2013a. « Rate-distortion analysis of dead-zone plus uniform threshold scalar quantization and its application-part I: fundamental theory. ». *IEEE transactions on image processing,* vol. 22, p. 202-14.

Sun, J., and others. 2013b. « Rate-Distortion Analysis of Dead-Zone Plus Uniform Threshold Scalar Quantization and Its Application-Part II: Two-Pass VBR Coding for H.264/AVC. ». *IEEE transactions on image processing,* vol. 22, p. 215-28.

Tourapis, A. 2002. « Enhanced Predictive Zonal Search for Single and Multiple Frame Motion Estimation ». *SPIE Proceedings, Visual Communications and Image Processing* vol. 4671, p. 1069-1079.

Tourapis, A., O. Au, and M. Liou. 2001. « Predictive Motion Vector Field Adaptive Search Technique (PMVFAST) - Enhancing Block Based Motion Estimation ». *SPIE*

Proceedings. Visual Communications and Image Processing 2001, vol. 4310, p. 883--892.

Wang, J., X. Yu, and D. He. 2011. « Hard-decision Quantization with Adaptive Reconstruction Levels for High Efficiency Video Coding ». *12th Canadian Workshop on Information Theory (CWIT)*, p. 62-65.

Wang, S., and others. 2011. « Rate-SSIM optimization for video coding ». *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 833-836.

Wang, S., and others. 2012. « SSIM-Motivated Rate Distortion Optimization for Video Coding ». *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 22, n° 4, p. 516 - 529.

Wang, Z. and A.C. Bovik. 2006. *Modern Image Quality Assessment*. Coll. « Synthesis lectures on image, video & multimedia processing ». Morgan & Claypool.

Wang, Z., and others. 2004. « Image Quality Assessment: From Error Visibility to Structural Similarity ». *IEEE Transactions on Image Processing,* vol. 13, p. 600-612.

Wang, Z. and Q. Li. 2007. « Video Quality Assessment Using a Statistical Model of Human Visual Speed Perception ». *Journal of the Optical Society of America,* vol. 24, p. 61-69.

Wiegand, T. and B. Girod. 2001. « Lagrange multiplier selection in hybrid video coder control ». *Proceedings. 2001 International Conference on Image Processing,* vol. 3, p. 542-545.

Wiegand, T., and others. 2003. « Rate-constrained coder control and comparison of video coding standards ». *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 13, n° 7, p. 688-703.

Xie, J. and L.T. Chia. 2008. « Study on the distribution of DCT residues and its application to R-D analysis of video coding ». *Journal of Visual Communication and Image Representation archive,* vol. 19, n° 7, p. 411–425.

Yarow, J. 2013. « Apple's Surprisingly Steady iTunes Growth ». < http://www.businessinsider.com/chart-of-the-day-itunes-revenue-2013-5 (last accessed 07/18/2013) >.

« YUV test sequences ». 2010. < http://media.xiph.org/video/derf/ (last accessed 07/18/2013)
http://trace.eas.asu.edu/yuv/ (last accessed 07/18/2013)
http://www.videocoders.com/yuv.html (last accessed 07/18/2013) >.

Zhao, X., and others. 2010. « Novel Statistical Modeling, Analysis and Implementation of Rate-Distortion Estimation for H.264/AVC Coders ». *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 20, n^{o} 5, p. 647-660.