ABSTRACT

Nowadays, the demand for more video compression keeps growing following the increasing wide scale deployment of applications where video compression plays a key enabling role. However, the current video compression standards are mainly optimized for objective quality metrics like the mean squared error with a rather low correlation with the user’s subjective assessment. The challenge addressed in this Thesis regards the inclusion in the state-of-the-art H.264/AVC video compression standard of additional coding tools able to increase its rate-distortion performance by exploiting some of the human visual system features. This challenge has been addressed by introducing in the H.264/AVC basic architecture two additional perceptually driven modifications. First, a perceptual coefficients pruning method using the thresholds provided by a selected just noticeable distortion model to set some of the DCT coefficients to zero has been implemented; after, an adaptive quantization method to adjust the introduced quantization error based on the same thresholds has been implemented. The main conclusion of the work reported in this Thesis is that it is worthwhile to adopt a perceptual approach in the optimization of the H.264/AVC video codec, reaching bitrate reductions which may go up to 35% for the same perceptual quality.

Index Terms— H.264/AVC, perceptual video coding, just noticeable distortion, transform, coefficients pruning, quantization

1. INTRODUCTION

Following the increasing wide scale deployment of daily life applications where video compression plays a key enabling role, the demand for more video compression keeps growing. Video compression or video coding technologies have the target to efficiently represent digital video data to allow the easier transmission and storage of this type of data. Therefore, it implies a complex balance between video quality, coding bitrate, encoding and decoding complexities, robustness to data losses and errors, ease of editing, random access, end-to-end delay, and a number of other relevant factors depending on the target application scenario. The main challenge in video coding is, thus, to reduce the compressed video data size for a target video quality or maximizing the video quality for a target coding rate. To compress the video data, compression algorithms exploit the redundancy and irrelevance in the original digital data, through largely used tools such as the discrete cosine transform (DCT), motion estimation and compensation, quantization and entropy coding. These mechanisms have to consider the relevant application constraints in terms of the available channel bandwidth and buffer sizes which are limited for most application scenarios. Thus, the critical need to design rate control solutions to achieve the best visual quality under certain relevant constraints, deeply associated to the application [1].

Traditionally, video coding optimization has been made based on simple objective quality metrics such as the mean squared error (MSE) and the peak signal to noise ratio (PSNR). These metrics do not accurately express the perceived distortion/quality as assessed by human subjects through the human visual system (HVS) and more appropriate quality metrics may be used to maximize the subjective quality for a certain target bitrate.

Some HVS properties which have been recently exploited in terms of video compression are the texture masking, the intensity contrast masking, the spatial frequency sensitivity and the preservation of object boundaries effects [2]. Since the 90s, several video compression standards have been developed by the Video Coding Experts Group (VCEG) from the International Telecommunications Union (ITU), more precisely the ITU Telecommunication standardization sector (ITU-T) and the Motion Picture Experts Group (MPEG) from the International Standards Organization/ International Electrotechnical Commission (ISO/IEC). The most recent is the H.264/AVC (Advanced Video Coding) video compression standard that was developed with the purpose to increase by 50% the compression performance provided by all previously available video compression standards while also providing a “network-friendly” video representation. The target applications included conversational, e.g. videotelephony and videoconference, and non-conversational scenarios, e.g. storage, broadcast and streaming, using bitrates from 50 kbps to 8 Mbps and more [3].
In this context, the main objective of this work is to design, implement and evaluate perceptually driven coding tools to be integrated into a standard H.264/AVC video codec, targeting the improvement of the rate-distortion performance and of the subjective video quality for a target bitrate, ideally without increasing too much the encoding complexity. The improvement of the H.264/AVC video codec must be made through two perceptually driven coding tools based on a JND model, namely a DCT coefficients pruning method and a JND adaptive quantization method.

The rest of this paper is organized as follows. Section 2 provides a review on the perceptual video coding concepts and technologies in the literature, starting with a review of the HVS properties, the available video coding standards and, finally, the most relevant perceptual video coding solutions. Next, Section 3 and Section 4 present the perceptually driven coding tools implemented in this work, notably a DCT coefficients pruning mechanism targeting the elimination of non-perceptually relevant DCT coefficients using a just noticeable distortion (JND) model, and an adaptive quantization mechanism introducing error according to the same JND model. With this purpose, the architecture, walkthrough and metrics used are first presented; after, the performance results are presented and analyzed to derive the main conclusions associated to the performance of the implemented methods. Finally, Section 5 is reserved for the description of eventual further work.

2. REVIEWING PERCEPTUAL VIDEO CODING CONCEPTS AND TOOLS

The first objective of this section is to briefly review the HVS; after, the evolution of video coding standards and recommendations is reviewed; and, finally, some of the most relevant perceptual video coding solutions in the literature are presented.

2.1. Brief Overview on the Human Visual System

The HVS includes the eye and also a part of the brain which is dedicated to process the visual information, through memory and knowledge. The eye is structured into three layers and each one has a specific function; the most important elements are the photoreceptors which transform the luminous stimuli into nervous impulses [4][5].

To exploit the HVS properties, perceptual models are used. The study of visual perception begun with Helmholtz, who defended the unconscious interference. After, the Gestalt theory focused on the understanding of visual components as a collection or an organized joint [6]. Visual perception can also be studied through psychophysics, an important concept is the so-called just noticeable distortion (JND), which is a threshold defining the smallest detectable difference between a starting and secondary levels of a particular sensory stimulus. The main psychophysical methods are the Weber’s law and the Stevens’ law [7][8].

2.2. Brief Overview on Video Coding Standards

The video compression saga started with the ITU-T H.261 Recommendation which was designed to address bidirectional and unidirectional visual communications at data rates multiples of 64 kbit/s, between 40 kbit/s to 2000 kbit/s and had as main target applications videotelephony and videoconference [9]. Afterwards, the MPEG-1 Video standard was designed to compress Video Home System (VHS) quality raw digital video down to around 1.2 Mbit/s without excessive quality loss, so to allow digital movie storage in CD-ROMs [10]. In 1994, after a joint effort between the ITU-T and ISO/IEC, the H.262/MPEG-2 Video standard was developed targeting digital television and thus interlaced video at higher rates and spatial resolutions than MPEG-1 Video, notably up to High Definition (HD) [11]. Soon after, the ITU-T H.263 Recommendation was developed to optimize video quality for lower bitrates, especially targeting applications such as visual telephony in copper telephone lines and mobile networks. Meanwhile, ISO/IEC developed the MPEG-4 Visual standard, providing increased compression efficiency but also adopting an object-based representation framework with high flexibility and advanced interaction capabilities. The main MPEG-4 Visual standard target applications ranged from digital television to mobile and Internet video streaming video and games [12]. More recently, the H.264/AVC video compression standard was developed with the intent of increasing by 50% the compression performance provided by all previously available standards while also providing a “network-friendly” video representation [3]. The target applications are the aforementioned in section 1.

2.3. Reviewing Perceptual Video Coding Solutions

This section intends to present a brief review on the most relevant perceptual video coding solutions in the literature since this Thesis targets this type of video compression approach. The main selection criteria for the coding solutions to be reviewed in the following are their compression performance as well as the novel way they integrate HVS features and associated tools in the video codec. Reviewing these solutions is fundamental to have a basic understanding of the state-of-the-art on perceptual video coding and, thus, to better decide which coding path to follow after in this work.

In [13] and [14], Chen and Guillenm propose a H.264/AVC perceptual video coding solution based on a Fooved Just Noticeable Distortion (FJND) model [13]. The foveated model is developed to further exploit the perceptual redundancy focusing on the fact that the visual acuity decreases with an increased eccentricity. In this solution, bit allocation and rate-distortion optimization (RDO) algorithms are proposed based on the FJND model with the target to achieve a better visual quality for the same rate. The basic idea is to use a FJND model to further exploit the perceptual redundancy by controlling the H.264/AVC Quantization Parameter (QP).
In [15], Mak and Ngan propose a novel approach for incorporating a DCT domain JND model in a H.264/AVC encoder to reduce the bitrate without visual quality loss. The basic idea is to remove the transformed coefficients with a magnitude lower than the JND threshold. The HVS properties mainly considered are the texture masking, intensity contrast masking, spatial frequency sensitivity and preservation of object boundaries.

In [16], Mai, Yang, Kuang and Po propose a novel motion estimation method based on the Structural Similarity (SSIM) metric for H.264/AVC inter prediction. The basic idea is to use a perceptual metric like the SSIM metric in the ME process instead of the usual SAD, with the target to reduce the bitrate and encoding time while maintaining the same perceptual video quality.

In [17], Huang and Lin propose a novel 4D perceptual quantization model for H.264/AVC bitrate control (PQrc). This solution includes two major encoding modules:
- Perceptual frame-level bit allocation using a 1D temporal pattern, depicted as the energy transition table, which is used to predict the frame complexity and determine proper rate budgets;
- Macroblock-level quantizer decision using a 3D rate pattern, formed as the bit-complexity-quantization (BCQ) model, in which the tangent slope of a BCQ curve is unique information to find a proper MB quantizer.

The basic idea is to use a 4D temporal BCQ model, this means a 1D temporal pattern to predict the frame complexity and to determine proper budget bits, and a 3D rate pattern, depicted as a BCQ model, to reduce the bitrate while improving the SNR quality and control accuracy. The HVS properties exploited are basically the just noticeable distortions.

3. A JND MODEL BASED COEFFICIENTS PRUNING METHOD FOR H.264/AVC VIDEO CODING

This section presents the first perceptually driven modification made to the H.264/AVC video codec with the target to eliminate the transform coefficients which are perceptually irrelevant according to an adopted JND model. After describing the coefficients pruning solution, the performance results obtained in the context of the H.264/AVC JM 16.2 reference software are presented and analyzed.

3.1. The Perceptual Coefficients Pruning Method

The basic idea underpinning the adopted perceptual coefficients pruning method is to remove, by setting them to zero, all the transform coefficients which have a magnitude lower than the corresponding JND threshold determined using an appropriate JND model since they should be perceptually irrelevant.

3.1.1. Architecture and Walkthrough

The major changes associated to the novel tool regard the determination of the pruning thresholds using the selected JND model and the transform coefficients pruning process included before the coefficients quantization. It is important to note that the proposed codec modifications only refer to the encoder and they do not imply any change in the H.264/AVC syntax and semantics as well as at the decoder; this implies that still fully compliant H.264/AVC bit streams are created with the perceptually driven video codec.

![Figure 3.1 – Improved H.264/AVC codec architecture including the JND model and the coefficients pruning method.](image_url)
6. Transform: The residual block is transformed with the ICT; the output are the transformed coefficients of the JND thresholded prediction residual;
7. Coefficients pruning:
   a. If the absolute value of a prediction residual ICT coefficient is larger than the corresponding JND threshold, go to step 8;
   b. Otherwise, the prediction residual ICT coefficient is set to zero meaning that the corresponding coefficient is NOT perceptually relevant and, thus, may be pruned, saving the associated rate; in this context, pruning means setting the value to zero;
8. Quantization: The transform coefficients which ‘survived’ the perceptual pruning method are now quantized;
9. Entropy encoding: The quantized coefficients are entropy coded with CABAC.

**Decoding Path (also within the encoder):**

1. Scaling & Inv. Transform: The quantized ICT coefficients are scaled (Q−1) and inverse transformed (F−1) to produce the residual block (D′n);
2. Motion Compensation and Reconstruction: The motion compensated prediction block PRED is added to D′n to create a reconstructed block;
3. Deblocking Filter: A deblocking filter is applied to the previously reconstructed blocks to reduce the blocking effects and the decoded picture is created from a series of filtered blocks F′n.

### 3.1.4. Novel Tools Description

This section describes the novel tools required to perform the perceptual pruning of the transform coefficients, notably the JND thresholds determination and the perceptual coefficients pruning method. The adopted pruning solution is based on the perceptual codec reviewed in section 2.3 designed by Mak and Ngan [4].

**A) JND Threshold Determination**

The adopted JND thresholds determination process relies on a spatial JND model which exploits three human visual system masking aspects through appropriate sub-models:

- **Frequency Band decomposition masking model** which exploits the different sensitivity of the human eye to the noise introduced at different spatial frequencies. The model is constituted by the default perceptual matrices adopted in the H.264/AVC reference software:
  
  \[
  JND_{\text{band}}^{\text{intr}}(l,j) = \begin{bmatrix}
  6 & 13 & 20 & 28 \\
  13 & 20 & 28 & 32 \\
  20 & 28 & 32 & 42 \\
  28 & 32 & 42 & 52 \\
  \end{bmatrix}
  \]

  (1)

  \[
  JND_{\text{band}}^{\text{extr}}(l,j) = \begin{bmatrix}
  14 & 20 & 24 & 27 \\
  20 & 24 & 27 & 30 \\
  24 & 27 & 30 & 34 \\
  \end{bmatrix}
  \]

  (2)

- **Luminance variations masking model** which exploits the masking effect associated to luminance variations in different image regions. The model is based on the Weber-Fechner law and is defined by equation (3) where \( \bar{E}(k) \) denotes the average luminance intensity in block \( k \):

  \[
  JND_{\text{lam}}(k) = \begin{cases}
  0.048 \cdot \bar{E}(k) + 4 & \bar{E}(k) \leq 62 \\
  \frac{1}{2} & 62 < \bar{E}(k) < 115 \\
  0.021 \cdot \bar{E}(k) - 1.464 & \bar{E}(k) \geq 115 
  \end{cases}
  \]

  (3)

- **Pattern masking effects model** which exploits the presence of some patterns in the image. To exploit this masking effect, the Foley-Boynton model was adopted. The pattern masking model is defined through equation (4), where \( E(i,j,k) \) denotes the normalized contrast energy and \( c \) is 0.6.

  \[
  JND_{\text{pat}}(i,j,k) = \begin{cases}
  1 & \text{if } i,j = 0 \\
  \max(1, E(i,j,k)^c) & \text{otherwise}
  \end{cases}
  \]

  (4)

  The final JND threshold \( J_T(i,j,k) \) is computed as:

  \[
  J_T(i,j,k) = JND_{\text{band}}(i,j) \cdot JND_{\text{lam}}(k) \cdot JND_{\text{pat}}(i,j,k)
  \]

  (5)

**B) Perceptual Coefficients Pruning Method**

The perceptual coefficients pruning method consists in setting to zero all the transform coefficients which have an absolute magnitude lower that the corresponding JND threshold given by the adopted JND model described above. The decision to prune an ICT coefficient is based on equation (6), where \( Y_p \) represents the pruned transform coefficients, \( Y \) stands for the transform coefficients and \( J_T \) represents the relevant JND threshold as defined by equation (5).

  \[
  Y_p(i,j,k) = \begin{cases}
  Y(i,j,k) & \text{if } |Y(i,j,k)| > J_T(i,j,k) \\
  0 & \text{otherwise}
  \end{cases}
  \]

  (6)

  To apply this pruning tool, it is important to know with which DCT, 4×4 or 8×8, the signal was transformed, and apply the pruning process not only to the luminance coefficients but also to the chrominance coefficients. The JND model computes different thresholds for the 4×4 and the 8×8 ICT blocks.

  To assess the pruning method performance, two specific statistical metrics are used:

  - **Average number of zeroed coefficients at MB level due to the perceptual coefficients pruning method**

    This statistical metric effectively measures the coefficients which were set to zero due to the additional perceptually driven coefficients pruning method. This statistical metric is defined in equation (7), where \( \text{Avg. coeff. MB} \), defined in equation (8), is the average number of coefficients zeroed at MB level after the quantization in each frame, \( \text{nr. coef. zero} \) is the number of coefficients zeroed in the frame and \( \text{nr. MB coded} \) is the number of MBs in a frame with coded coefficients. This metric is computed over all the frames in the sequence, and then its average in time is computed as expressed in equation (9).

    \[
    \text{Avg. zeroed coef. MB} = \frac{\text{Avg. coeff. MB}}{\text{nr. MB coded}}
    \]

    (7)

    \[
    \text{Avg. coeff. MB} = \frac{\sum_{\text{frame}} \text{nr. coef. zero}}{\text{nr. MB coded}}
    \]

    (8)

    \[
    \text{nr. MB coded} = \frac{\sum_{\text{frame}} \text{nr. MB coded}}{\text{nr. frame}}
    \]

    (9)
This statistical metric computes the average zigzag position in the 4×4 blocks corresponding to the zeroed coefficients due to the perceptual coefficients pruning method (position Є [1, 16]). This statistical metric provides an idea on the bandwidth zone where the coefficients are being zeroed. The statistical metric is computed with equation (10) where $nr_{coef\_zero}$ represents the number of zeroed coefficients in the position $i$ of each 4×4 block for a frame; this statistical metric is computed in the universe of the MBs coded with the 4×4 DCT.

$$Avg_{zigzag\_position\_MB} = \frac{\sum_{i=0}^{nr_{frames}} Avg_{zigzag\_position\_MB\_i}}{nr_{frames}}$$

$$Avg_{zigzag\_position\_sequence} = \frac{\sum_{i=0}^{nr_{frames}} Avg_{zigzag\_position\_MB\_i}}{nr_{frames}}$$

3.2. Performance Evaluation

3.2.1. Test Conditions

For the test experiments, the H.264/AVC reference software, version JM 16.2 (FRExt), has been used, notably the JND model and the perceptual coefficients pruning method described in the previous section were implemented in the context of this reference software. Further test conditions include [19]:

- **GOP prediction structure:** IBPPBPPBBP..., with a single Intra frame at the beginning.
- **Rate control:** RDO in the high complexity mode.
- **Test sequences and resolutions:**
  - Foreman and Mobile
    - Spatial resolution: CIF
    - Frame rate: 30 fps
    - Total number of frames: 300 frames
  - Panslow and Spincalendar
    - Spatial resolution: 1280x720
    - Frame rate: 60 fps
    - Total number of frames: 150 frames
  - Playing_cards and Toys_and_calendar
    - Spatial resolution: 1920x1080
    - Frame rate: 24 fps
    - Total number of frames: 60 frames
- **Quantization parameters:** Several QP values groups were used; To simplify, each group is presented as $G_x = (QPI, QPP, QPB)$ where $G_x$ represents a group x and $QPy$ the quantization parameter when y are the I, P and B frames:
  - G1 = (12, 12, 12)
  - G2 = (16, 16, 16)
  - G3 = (22, 23, 24)

3.2.2. Results and Analysis

To assess the RD performance, RD charts for the PSNR and MS-SSIM objective quality metrics versus the bitrate were obtained for each test sequence and for each adopted RD point; in the following, only the most relevant charts will be presented.

- **RD Performance: PSNR versus Bitrate**

As expected, the PSNR increases with the rate, first quite quickly and after with a rather linear variation; in this case, the various rates were obtained from the various quantization parameters combinations, corresponding to the $G_x$ labels in the charts. For low and medium sequences, there are no significant variations in terms of RD performance between the two codecs (HP and JND_CP) under comparison. For the higher resolution sequences, there are evident RD performance gains obtained with the H.264/AVC JND_CP solution, notably for the higher rates. The rate gains go up to about 8% (7 Mbit/s) for the same PSNR for the last RD point or PSNR gains of about 1.1% (0.5dB), for the same rate, for the last RD point. The RD gains are larger for the higher resolutions because in high resolution sequences each MB corresponds to a tinnier physical area; in this context, the redundancy is higher and, therefore, the coefficients are lower and consequently the coefficients pruned have lower impact in the video quality.
RD Performance: MS-SSIM versus Bitrate

As expected, the MS-SSIM increases with the rate, first rather quickly and after saturating the quality for the higher bitrates; this basically means that the subjective quality saturates when the rate increases above a certain value implying that, contrary to what is said by the PSNR, the subjective quality does not continuously increase with the rate since non-perceptible details are being sent at some stage. For the low and medium resolution sequences, the H.264/AVC JND_CP solution shows rate gains up to about 8%, for the same MS-SSIM, for the last RD point. For the higher resolution sequence, there are more evident RD performance gains associated to the H.264/AVC JND_CP solution, notably for the higher rates. The rate gains go up to about 13% (11 Mbit/s), for the same MS-SSIM, for the last RD point.

Figure 3.3 - MS-SSIM RD Performance of the Playing_cards sequence

Comparing the MS-SSIM RD performance with the PSNR RD performance, the following conclusion may be taken: a high quality level in terms of MS-SSIM can be achieved with a lower bitrate than in terms of PSNR. This means that PSNR improvements after a certain rate are non-perceptible since the subjective quality saturates due to the HVS perception limitations. The recognition of this effect may allow adopting lower coding rates while still achieving the same high subjective quality.

Average number of zeroed coefficients per MB due to the perceptual coefficients pruning method

For lower's Gs, this means higher rates and qualities, the average number of zeroed coefficients is larger because for the higher rates the reference software uses lower thresholds (i.e., has less coefficients to set to zero). Therefore, when the perceptual coefficients pruning method is applied, there will be more coefficients under the JND threshold. The highest average number of zeroed coefficients in a MB is around 10.2 coefficients for the lowest G in the Playing_cards sequence. On the contrary, for the higher Gs the average number of zeroed coefficients per MB is close to zero since most of the coefficients which are perceptually irrelevant are already set to zero by the quantization process. The average number of zeroed coefficients per MB increases as the resolution increases; this happens because in high resolution sequences each MB corresponds to a tinier physical area; in this context, the redundancy is higher and, therefore, the coefficients are lower and consequently there are more coefficients pruned. The sequences with strong variations in time, pattern areas or a low luminance level have a higher average number of zeroed coefficients because the frequency band decomposition masking model exploits the type and quantity of variations in each image, the pattern masking model exploits the contrast and the luminance masking model used in the JND model provides a higher JND threshold for sequences with a darker background (low level of luminance), respectively.

Average zigzag position of the zeroed coefficients exclusively due to the perceptual coefficients pruning method

The average zigzag position for the zeroed coefficients in 4×4 blocks increases with the rate. This is expected as the higher is the rate, the lower is the quantization step and, thus, the more irrelevant coefficients are coded, if not filtered by the pruning method. For the first Gs, this means higher rates and qualities, the average zigzag position is slightly higher, going up to 11.4, on average, for G1 in the Mobile sequence. This happens since the reference software has a lower threshold for the higher bitrates (i.e. the quantization only sets to zero coefficients with very high frequency); consequently, when the perceptual coefficients pruning method is applied, since the human eye is less sensitive to high frequencies, the coefficients set to zero will be in a frequency range a little bit higher comparing to the other Gs. On average, the zigzag position where it is more likely for coefficients to be set to zero is, approximately, 10 with variance, approximately, 2 for low and medium resolution and, approximately, 9 with variance, approximately, 2 for high resolution. These positions correspond to the middle range frequencies. This happens since the frequency band decomposition masking model applies a higher threshold for the higher frequencies; thus, in these frequencies more coefficients are set to zero. Therefore, it could be assumed that this method has a bigger impact on the high frequencies; however, since the quantization sets coefficients to zero in the higher frequencies, this method ends by having a bigger impact in the middle frequencies. The sequences with stronger variations in time show a higher value for the average zigzag position of the zeroed coefficients.

3.2.3. Conclusions

The main conclusion of this section is that the adopted perceptual coefficients pruning method may have some positive impact on the RD performance, notably for sequences with lower luminance levels (e.g. Playing_cards vs Toys_and_calendar) and objects with patterns (e.g. Panslow vs Spincalendar), especially for the low QP values and higher resolutions. The H.264/AVC JND_CP codec can achieve PSNR RD gains that are only evident for the higher resolutions and for the higher rates. The rate gain goes up to
8% for the last RD point and the PSNR gain goes up to 0.5 dB for the last RD point. Also, the MS-SSIM RD performance can achieve rate gains up to 8% for low and medium resolutions and 13% for high resolutions, both for the last RD point. A good quality in terms of MS-SSIM can be achieved with a lower rate (around 700 kbit/s for Foreman sequence) than in terms of PSNR where a good quality is only achieved with a higher rate (larger than 3.7 Mbit/s for the Foreman sequence).

Regarding the number of zeroed coefficients due to the perceptual coefficients pruning method, it can go up to around 9.7 coefficients, on average, for the Playing_cards sequence in the lowest G. These zeroed coefficients are more likely to be located in between the 8th and 12th zigzag position in a 4×4 block. Therefore, the coefficients zeroed by the perceptual coefficients pruning method are typically in the middle to high frequencies.

4. A JND MODEL BASED ADAPTIVE QUANTIZATION METHOD FOR H.264/AVC VIDEO CODING

This section presents the second perceptually driven modification to the reference H.264/AVC video codec with the objective of quantizing the ICT coefficients based on an adopted JND model. After describing the new ICT quantization solution, the performance results obtained in the context of the H.264/AVC JM 16.2 reference software using several objective quality metrics are presented and analyzed.

4.1. The Adaptive Quantization Method

The basic idea of the new perceptually driven quantization method for the ICT coefficients is to adjust the QP values based on the JND thresholds computed based on a selected JND model. With this purpose, a distortion weight will be determined for each MB to be applied to each initially assigned QP value in order to get a new JND adapted QP value. The basic idea is to code the ICT coefficients with an accuracy controlled associated to their perceptual relevance.

4.1.2. Architecture and Walkthrough

The improved encoder architecture already including the JND adaptive quantization related tools is presented in Figure 4.1 while Figure 4.2 presents the improved encoder architecture including also the ICT perceptual coefficients pruning solution presented in the previous section. The major architectural change regards the control of the QP value based on the JND thresholds. It is important to note that the proposed codec modifications only refer to the encoder and they do not imply any change in the H.264/AVC syntax and semantics, meaning that still fully compliant H.264/AVC bit streams are created.

The walkthrough of the new perceptual video codec with special emphasis on the novel modules related to the QP perceptual control based on the JND model, which are listed in bold, is presented below. The 7th step is only included for the improved perceptual codec using both the JND related methods presented in this paper, this means coefficients pruning and JND adaptive quantization.

**Forward/Encoding Path**
1. MB Division: As presented in Section 3.1.3;
2. JND thresholds determination: As presented in Section 3.1.3;
3. Motion Estimation: As presented in Section 3.1.3;
4. MB prediction: As presented in Section 3.1.3;
5. Residue Computation: As presented in Section 3.1.3;
6. Transform: As presented in Section 3.1.3;
7. Coefficients Pruning: As presented in Section 3.1.3;
8. JND QP Adaptation: Adapt the QP value for each MB based on the JND thresholds as defined in the next section;
9. Quantization: Quantize the transformed (and eventually pruned) coefficients with the JND adapted QP;
10. Entropy encoding: As presented in Section 3.1.3;

**Decoding Path (also within the encoder):** The decoder is the same as presented in Section 3.1.3 since there are no changes in the decoder. This also reflects the fact that the
proposed JND related tools do not impact the H.264/AVC compliance, and thus the same (normative) decoder is used.

4.1.4. Novel Tools Description

This section describes the novel tool required to perform the QP perceptual adaptation, notably the computation of the new QP. This solution is based on the perceptual video coding solutions presented in Section 2.3 developed by Chen and Guillemt [13][14].

The JND adaptive quantization method consists in adapting the QP value for each MB, taking into account its perceptual relevance. The basic idea is to use the JND thresholds to determine a new QP value, if the average value of the JND thresholds in a MB is higher than the average value of the JND threshold in a frame (using the thresholds for the relevant coding mode, e.g. 4x4 DCT or 8x8 DCT and INTRA or INTER modes) meaning that the MB is perceptually less relevant regarding the average relevance of the frame. Consequently, the new QP will be higher than the QP determined by the H.264/AVC JM reference software, exploiting the HVS behavior to mask some additional quantization noise, thus saving some bitrate. This tool modifies the QP value as initially determined by the H.264/AVC JM reference software by using a weighted distortion \( dist \) computed based on the JND thresholds computed using the JND model presented in Section 3.1.4. The QP value is adapted as follows:

\[
Q_{\text{JND}} = Q_P + dist, \quad (11)
\]

where \( Q_P \) is the initially determined QP value and \( Q_{\text{JND}} \) is the adapted QP value.

The weighted distortion \( dist \) is computed through equation (12) where \( \text{avg } \text{JND}_{\text{MB}} \) is the average value of the JND thresholds for MBs, \( \text{avg } \text{JND}_{\text{frame}} \) is the average value of the JND thresholds for the frame, computed by equation (14) where \( J_T \) is the JND threshold for each coefficient in the MB.

\[
dist_i = 0.7 + 0.6 \times \frac{1}{1 + \exp\left(-4 \times \frac{\text{avg } \text{JND}_{\text{MB}}}{\text{avg } \text{JND}_{\text{frame}}} - \frac{\text{avg } \text{JND}_{\text{frame}}}{16} \right)} \quad (12)
\]

\[
\text{avg } \text{JND}_{\text{MB}} = \frac{\sum_{i=0}^{\text{height}} \sum_{j=0}^{\text{width}} \text{J}_T[i][j] \text{J}_T[i][j]}{16 \times 16} \quad (13)
\]

\[
\text{avg } \text{JND}_{\text{frame}} = \frac{\sum_{i=0}^{\text{height}-1} \sum_{j=0}^{\text{width}-1} \text{J}_T[i][j]}{\text{height} \times \text{width}} \quad (14)
\]

In summary, the average JND threshold for the frame and for each MB are computed using the JND model presented in Section 3.1.4 A. Afterwards, the weighted distortion for each MB is computed and the new QP value is determined by this distortion as in (11).

To avoid a subjectively negative flickering effect due to the variation of the QP value between MBs, the QP variations are limited: the QP value can only decrease 1 and increase until 3 as defined by equation (15). After the determination of \( Q_{\text{JND}} \) with equation (11), the conditions in (15) are checked: if \( dist_i \) is lower than 1 and \( Q_{\text{JND}} \) is less than \( Q_P - 1 \) or if \( dist_i \) is higher than 1 and \( Q_{\text{JND}} \) is higher than \( Q_P + 3 \), \( Q_{\text{JND}} \) will be further changed: in the first case, \( Q_{\text{JND}} \) will be set to \( Q_P - 1 \) while in the second case \( Q_{\text{JND}} \) will be set to \( Q_P + 3 \).

\[
\left\{ \begin{array}{l}
Q_{\text{JND}} = Q_P + 3 & \text{if } (dist_i > 1) \& \& (Q_{\text{JND}} - Q_P > 3) \\
Q_{\text{JND}} = Q_P - 1 & \text{if } (dist_i < 1) \& \& (Q_{\text{JND}} - Q_P) > 1
\end{array} \right. \quad (15)
\]

4.2. Performance Evaluation

This section intends to assess the performance of the presented perceptual H.264/AVC codec, including the proposed adaptive quantization method. With this purpose in mind, first the adopted test conditions are presented, including the selected performance metrics and the benchmarks; after, the performance results are presented and analyzed.

4.2.1. Test Conditions

For the tests, the H.264/AVC reference software, version JM 16.2 (FRExt), has been used, notably the High profile for the reasons mentioned before. Thus, the adopted JND model, the coefficients pruning method described in the previous section and the JND adaptive quantization method were implemented in the context of this reference software. Further test conditions included [19]:

- **GOP prediction structure**: As presented in Section 3.2.1.
- **Rate control**: As presented in Section 3.2.1.
- **Test sequences and resolutions**: As presented in Section 3.2.1.
- **Quantization parameters**: As presented in Section 3.2.1.
- **Coding benchmarks**: The proposed perceptual video codec with the JND adaptive QP and the proposed perceptual video codec including coefficients pruning and JND adaptive QP are compared with the H.264/AVC High profile codec and with the H.264/AVC based perceptual codec only with coefficients pruning. In the following, these codecs will be labeled as:
  - **HP** - H.264/AVC High profile codec
  - **JND_CP** - H.264/AVC based perceptual codec with coefficients pruning
  - **JND_QP** - H.264/AVC based perceptual codec with JND adaptive QP
  - **CP+JND_QP** - H.264/AVC based perceptual codec including coefficients pruning and JND adaptive QP

- **Performance metrics**: Besides the objective quality metrics already adopted in the previous section, another objective quality metric is adopted to perform a more complete RD performance assessment: VQM and RP compensated VQM. The second metric is needed because typically the VQM is not powerful enough to compare two codecs in terms of RD performance [20].

4.2.2. Results and Analysis

To evaluate the RD performance of the JND_QP and JND_CP+QP codecs four objective quality metrics are used
in this section: PSNR, MS-SSIM, VQM and RP compensated VQM; results were obtained for each RD point and each test sequence, but only the most relevant charts will be presented in the following.

- **RD Performance: PSNR versus Bitrate**
  
  From the PSNR results, it may be concluded that there are only losses relatively to the HP and JND_CP codecs. This does not come as a surprise as the PSNR relies on the mathematical difference between the luminances of the original and decoded sequences and, thus, does not necessarily accurately model the subjective quality. In fact, the JND adaptive QP method introduces additional quantization error for some MBs under the assumption that this additional error is not perceptible although the PSNR will still ‘complain’ about this additional error as the PSNR is still (mathematically) sensitive to this error; thus, it is simply normal that the PSNR RD performance for the JND_QP and the JND_CP+QP codecs is worst than for the HP and JND_CP codecs.

- **RD Performance: MSqSSIM versus Bitrate**
  
  The JND_QP and JND_CP+QP codecs present a maximum quality similar or inferior to the reference HP codec quality because the MS-SSIM, as aforementioned, is based on structure. Knowing that both the perceptual codecs modify the QP value for each MB, it is possible that adjacent MBs have a significant difference in the QP value (e.g. if the QP value as initially assigned by the H.264/AVC JM software reference is 33, the QP for the MB may be between a minimum value of 32 and a maximum value of 36); this does not happen for the HP codec since its RD performance was evaluated for constant QP values. Consequently, the MS-SSIM interprets this quality difference as block artifacts, and the quality metric is lower.

- **RD Performance: VQM versus Bitrate**
  
  The VQM decreases with the rate (thus the quality increases), first rather quickly and after tends to a value slightly higher than zero for the higher bitrates; this basically means that there is a good subjective quality when the rate increases above a certain value and the VQM is near to zero meaning the distortion is very low. The various rates are obtained from the various quantization parameters combinations corresponding to the Gx labels in the charts. The JND_QP and JND_CP+QP codecs present rate gains for all resolutions, especially for the highest rates; however, there are no VQM gains as the maximum subjective quality is the same for all codecs under test. For low resolution sequences and quantization steps between G2 and G5, there are some VQM losses; on the contrary, for high resolution sequences, there are some VQM gains.

- **RD Performance: RP compensated VQM versus Bitrate**
  
  For both the JND_QP and JND_CP+QP codecs and for all resolutions, there are rate gains and RP compensated VQM gains. The JND_QP codec can achieve RP compensated VQM RD gains through rate gains up to 32.5%/27% for the last RD point regarding the HP and JND_CP codecs, respectively, and RP compensated VQM gains up to 0.03/0.02 for the second to last RD point relatively to the HP and the JND_CP codecs, respectively. The JND_CP+QP codec can achieve RP compensated VQM RD gains through rate gains up to 35%/29.7% for the last RD point or VQM gains up to 0.04/0.02 for the second to last RD point for the HP and JND_CP codecs, respectively.

4.2.3. Conclusions

The main conclusions of this section are that the PSNR and the MS-SSIM quality metrics are not able to adequately express the subjective RD performance gains obtained with the proposed JND_QP and JND_CP+QP codecs because they are not designed to efficiently measure the subjective quality and, thus, have a low correlation with subjective quality scores. However, the effective assessment of the RD performance gains was possible with the VQM objective quality metric and its RP compensated version.
VQM RD gains for the JND_QP and JND_CP+QP codecs are mainly rate gains; however, there are also some VQM gains, especially for the high resolution sequences, going up to 0.02 and 0.03 when comparing with the HP codec and 0.01/0.02 when comparing with the JND_CP codec for the JND_QP and JND_CP+QP codecs, respectively. Regarding the RP compensated VQM RD gains for the JND_QP and JND_CP+QP codecs, there are both rate and RP compensated VQM gains. The highest RD gains for the JND_CP+QP codec are RP compensated VQM gains which go up to 0.02 and rate gains which go up to 35% for the low resolutions. Still for the same codec, the RP compensated VQM gains go up to 0.04 and the rate gains go up to 31% for the medium and high resolutions.

5. FUTURE WORK

Despite the encouraging results achieved with the combination of the two perceptually driven tools described in this paper, the developed video coding solutions still leaves room for improvements. An important module in the H.264/AVC that was not improved but it is a good candidate for perceptually related improvements is the motion estimation. The motion estimation module may be improved using the same basic ideas that were used to improve the transform and quantization processes following the computation of the JND model. In this context, a method similar to the one presented in [16] may be developed; the basic idea is to compute the distortion metric comparing the original and the prediction block to determine the prediction error using only the perceptually relevant residuals based on some filtering using the relevant JND thresholds.

6. ACKNOWLEDGEMENTS

First and foremost I would like to thank my supervisor in this thesis, Prof. Fernando Pereira for the valuable guidance, advice, devotion and dedication to this work. Besides, I would like to thank all the IT Image Group for their support, especially Matteo Naccari and Catarina Brites, for the availability and help on technical issues when I needed the most. Finally, an honorable mention goes to my family and friends for their understanding and support on me in completing this project. Without helps of the particular that mentioned above, I would have faced many difficulties while doing this project. A deep acknowledgement to all the above mentioned as well as to everyone close to me, which, in their own way, contributed for the development of this Thesis.

7. REFERENCES

http://brainconnection.posit science.com/topics/?main=anat/visi on-anat2
http://en.wikipedia.org/wiki/Stevens' power law