PERCEPTUAL WATERMARKING USING JUST NOTICEABLE DIFFERENCE MODEL BASED ON BLOCK CLASSIFICATION

Ivan Damnjanovic
Queen Mary, University of London
Mile End Road
London E14NS, UK

ivan.damnjanovic@elec.qmul.ac.uk

Ebroul Izquierdo
Queen Mary, University of London
Mile End Road
London E14NS, UK

ebroul.izquierdo@elec.qmul.ac.uk

ABSTRACT
One of the main goals of watermarking is to optimize capacity while preserving high video fidelity. The perceptual adjustment of the watermark is mainly based on Watson Just Noticeable Difference (JND) model. Recently, it was proposed to improve Watson model using the classification blocks inherent to the encoder in a compressed stream. Although, the new model outperforms the previous one, especially in increasing the watermark power in textured blocks, it still underestimates JNDs at block's edges. This work presents a detailed comparison of these two models and proposes a new method that exploits the good characteristics of the two available models. In addition, experimental results on perceptibility are reported.

Categories and Subject Descriptors
H.5.1 [**Information Systems**]: Multimedia Information Systems – *video, watermarking.*

General Terms
Algorithms, Measurement, Performance, Design, Security.

Keywords
Digital watermarking, MPEG2, perceptual adjustment, Just Noticeable Difference, block classification.

1. INTRODUCTION
The huge expansion of digital production and processing technology and the Internet have made possible to distribute and share unlimited quantities of digital material by anyone, anytime and anywhere. Digital watermarking arose as a possible solution to not only inherent copyright issues, but also a range of other interesting applications such as authentication, broadcast monitoring and data embedding.

Looking at real-time applications often required in Digital Video

Broadcasting or Internet and mobile streaming, several techniques have been reported in the literature aiming at watermarking in the compressed domain. Many of these are based on embedding a watermark into a video sequence using the spread spectrum paradigm. In their pioneering work, Hartung and Girod proposed to extend their technique for spread spectrum watermarking in uncompressed domain to compressed domain [1]. This is achieved by transforming a watermark signal into DCT values and then adding it to the DCT coefficients of the video sequence.

The power of the watermarking signal is bounded by perceptual visibility. Thus, the watermark must be embedded in such a way that it does not introduce visual artefacts to the host signal. Perceptual watermarking models are often based on Watson visual model developed for JPEG compression quantization matrices [2], [3]. The model estimates the perceptibility of changing coefficients in 8x8 DCT blocks. It consists of a sensitivity function and two masking components based on luminance and contrast masking. Following this approach, Zhang et al proposed an improved estimation of the JND in terms of luminance and contrast masking [4].

The aim of this work is twofold: To conduct a thorough comparison of these two models by assessing their performance in terms of capacity under the same imperceptibility conditions; and to derive a new better perceptual schema capitalizing on the good features of the two previous models. The performed analysis is constrained to compressed domain watermarking, specifically MPEG-2 video streams. Results of selected computer experiments are also reported.

2. WATERMARKING OF MPEG-2 VIDEO SEQUENCES
The targeted application is data embedding and indexing in professional environments where intentional attacks are not expected, so the watermark needs to be robust against typical video editing processes, such as transcoding, logo insertion, cross-fade etc. Hence, focus was given to requirements related to high imperceptibility and the trade off between imperceptibility and watermark capacity.

A minimum duration of the watermarking video segment from which it will be possible to extract the watermark is often limited by a time window of 5 seconds [5]. For the MPEG2 standard in PAL sequences it can be seen as 8 I frames. In this way only 8 frames needs to be processed at the watermark decoder. Due to temporal compression, the embedding space in inter-frames is considerably low, so this can be seen as reasonable trade-off between the capacity and the computational cost.
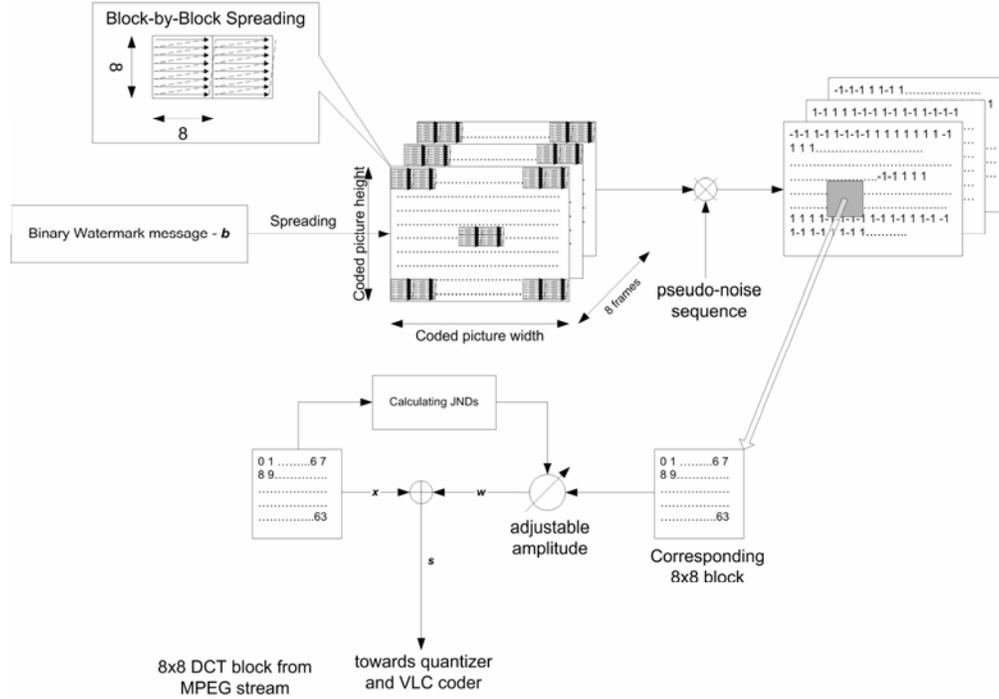
Figure 1. Watermarking embedding scheme

The principle of the watermarking scheme used in this work, is given in Figure 1. The scheme is based on the popular spread-spectrum paradigm where each of n watermarking bits is repeated 64 times to form 8x8 block. These blocks are then randomly spread through 8 watermarking frames and then modulated by pseudo sequence [7]. In that way, every watermark bit has almost the same Signal-to-Noise ratio, since the bits are evenly spread through textured, edge and plain areas. Before embedding the watermark to DCT coefficient, its amplitude is adjusted using the information from corresponding DCT block in the original sequence.

In the given watermarking system we are restricted to BPSK (Binary Phase Shift Keying) signalling, since watermarking bits are taking only two values either +1 or -1. In that way, since the input signal is not ideal for the given AWGN channel, mutual information between the input and the output signal will not be maximal leading to a lower maximum achievable capacity:

$$C_{BPSK} = 1 - \frac{1}{\sqrt{2\pi}} \int_{-5}^{+20} e^{-\frac{1}{2}\left(y - \sqrt{\frac{S}{N}}\right)^2} \cdot \log_2\left(1 + e^{-2y\sqrt{\frac{S}{N}}}\right) dy \quad (1)$$

Equation 1 is solvable numerically and represents the actual capacity boundary of our watermarking system. It was also shown [8] that with BPSK signalling it is possible to achieve bit error rate of 10-5 with an SNR of 9.6dB.

3. PERCEPTUAL WATERMARK AMPLITUDE ADJUSTMENT

One of the main watermarking requirements in a professional environment is high imperceptibility of the watermark and induced distortions. The watermarked material has to preserve video quality, i.e. the watermarked video must be indistinguishable from the original.

The JND models used in this work tends to exploit three basic types of phenomena: non-uniform frequency response of human eye (contrast sensitivity - t_{CSF}), sensitivity to the different brightness levels (luminance masking - t_l) and sensitivity to one frequency component in the presence of another (contrast or texture masking - t_c):

$$t_{JND}(n_1, n_2, i, j) = t_{CSF}(i, j) \times t_l(n_1, n_2, i, j) \times t_c(n_1, n_2, i, j) \quad (2)$$

where the indices n1 and n2 show the position of the 8x8 DCT block in the image or the video frame, while i and j represent position of the coefficient in the DCT-block.

The visibility threshold t_{CSF}, as a function of spatial frequency response in specific viewing conditions, is usually derived by the model presented in [6]. These thresholds for pre-determined viewing conditions can be also given in 8x8 contrast sensitivity table as the one used in our experiments and reproduced from [3].

The human visual system's sensitivity to variations in luminance is dependent on the local mean luminance. In the Watson model luminance adaptation is based on a power version of Weber-Fechner's law:

$$t_l^{WAT}(n_1,n_2,i,j) = \left(\frac{C(n_1,n_2,0,0)}{\overline{C(0,0)}} \right)^a \qquad (3)$$

where $\overline{C(0,0)}$ is the mean luminance of the image and a is the parameter that controls the effect of luminance masking on the model and its suggested value is 0.649.

Zhang et al. in [4] argued that Watson model over-simplifies the viewing conditions for practical images. They stated that gamma-correction of the display tube and ambient illumination falling on the display partially compensate effect of Weber-Fechner's law and as a result give higher visibility thresholds in either very dark or very bright regions, which Watson model fails to acknowledge. Hence, they approximate luminance thresholds with two functions, for low region (L≤128) and for high region of luminance (L>128):

$$t_l^{ZHA}(n_1,n_2,i,j) = \begin{cases} k_1 \left(1 - \frac{C(n_1,n_2,0,0)}{128} \right)^{\lambda_1} + 1 & if \quad C(n_1,n_2,0,0) \le 128 \\ k_2 \left(\frac{C(n_1,n_2,0,0)}{128} - 1 \right)^{\lambda_2} + 1 & otherwise \end{cases} \quad (4)$$

where k_1=2, k_2=0.8, λ_1=3, λ_2=2.

To incorporate contrast masking effect, the Watson model considers masking within a block and a particular DCT coefficient, since the contrast masking effect is strongest when two components have the same frequency, orientation and direction. In this model, the contrast masking threshold t_c is a function of DCT coefficient $C(n_1,n_2,i,j)$ and thresholds t_{CSF} and t_l :

$$t_C^{WAT}(n_1,n_2,i,j) = \max \left\{ 1 \quad , \quad \left(\frac{|C(n_1,n_2,i,j)|}{t_{CSF}(n_1,n_2,i,j) \cdot t_l(n_1,n_2,i,j)} \right)^w \right\} \quad (5)$$

where w is a constant that lies between 0 and 1, which may differ for each frequency, but usually have a constant value.

To evaluate the effect of contrast masking more accurately, it is essential to classify the DCT blocks according to their energy. It is a well known fact that noise is less visible in the regions where texture energy is high and it is easy to spot in smooth areas. Furthermore, the Human Visual System (HVS) is sensitive to the noise near a luminance edge in an image, since the edge structure is simpler than textured one.

DCT blocks are assigned in one of three classes: TEXTURE, EDGE and PLAIN using the algorithm given in [4]. According to the block class and its texture energy, the inter-band elevation factor is derived using the following formula:

$$\xi(n_1,n_2) = \begin{cases} 1 + 1.25 \cdot \frac{(E+H) - \mu_2}{2\mu_3 - \mu_2} & \text{for TEXTURE} \\ 1.25 & \text{for EDGE and } L+E > 400 \\ 1.125 & \text{for EDGE and } L+E \le 400 \\ 1 & \text{for PLAIN block} \end{cases} \quad (6)$$

where L, E and H are sums of AC coefficients in low, edge and high frequencies group and μ_2=290, μ_3=900.

To consider the intra-band masking effect Watson's contrast masking model was used:

$$t_C^{ZHA}(n_1,n_2,i,j) = \begin{cases} \xi(n_1,n_2) & \text{for } (i,j) \in L \cup E \\ & \text{in EDGE} \\ \xi(n_1,n_2) \cdot \max\left\{1, \left(\frac{|C(n_1,n_2,i,j)|}{t_{CSF}(n_1,n_2,i,j) \cdot t_l(n_1,n_2,i,j)} \right)^w \right\} & \text{otherwise} \end{cases} \quad (7)$$

AC coefficients in low and edge regions of EDGE blocks are excluded to avoid over-estimation of thresholds near edges.

In our experiments, we first attempted to find a suitable the value for w. In the Watson model this parameter was fixed to 0.7, which in specific set-up conditions overestimate the JND especially around edges (Figure 2). Zhang proposed a much lower value of w=0.36. On the other hand, Zhang model tends to underestimate JND in EDGE blocks and gives the low watermark power in edgy sequences. In addition, we observed that in the presence of noise around edges, Zhang method tends to augment it. Thus, noise becomes more visible and annoying. As a consequence of this observations, we propose a combined method that uses the Watson model with w=0.36 for EDGE and PLAIN blocks and Zhang improvements in TEXTURE blocks.

The proposed combined model, as it will be seen from experiments, outperforms both models by exploiting good characteristics of the Watson model in EDGE blocks and the Zhang model in Texture blocks. Concerning the image fidelity, it will be shown that it is still high and comparable with the result of the other two methods.
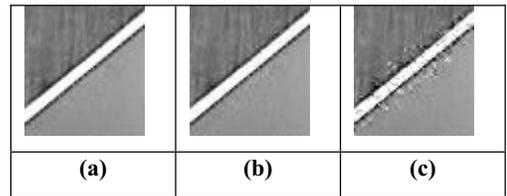

Figure 2. Edge detail from Table Tennis: (a) original and watermarked with Watson model (b) w=0.36 and (c) w=0.7

Table 1. Peak Signal to Noise Ratio comparison of three methods

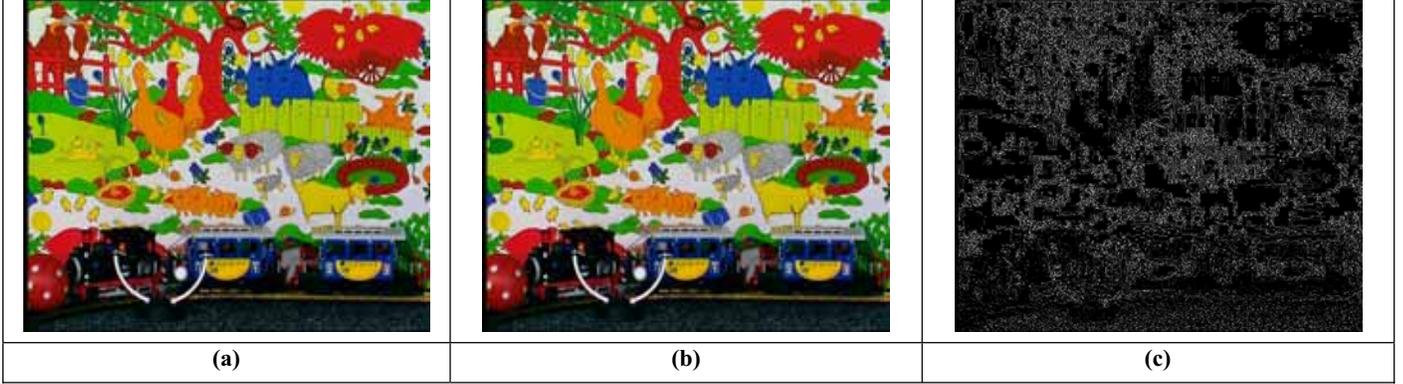|	Table Tennis			Flower Garden			Mobile and Calendar			Suzy			BBC3		
PSNR	Min	Avg	Max	Min	Avg	Max	Min	Avg	Max	Min	Avg	Max	Min	Avg	Max
Zhang	36.99	39.62	55.77	37.26	39.78	47.65	36.31	40.86	49.55	44.46	48.80	56.90	40.39	47.84	55.53
Watson	36.96	40.09	53.99	37.01	39.36	46.26	37.06	41.79	50.17	44.85	48.99	56.99	39.61	46.26	53.73
Proposed	36.79	39.58	53.10	36.54	38.88	45.82	35.82	40.05	47.64	44.15	48.31	55.61	38.98	45.53	53.01

Figure 3. Frame 345 from "Mobile and Calendar" sequence: original (a), watermarked with proposed model (b) and frame difference (c)

4. EPERIMENTAL RESULTS

The presented techniques were evaluated using five typical MPEG2 sequences (Table Tennis, Flower Garden, Mobile and Calendar, Suzy and BBC3). All sequences were 375 frames long, PAL (704x576, 25 fps), with GoP IBBP structure, size 12 and bit-rate 6 mbps.

The result of PSNR test is given in Table 1. A watermark is embedded in every of five sequences and they are compared with originals frame by frame. The table shows minimal PSNR values, maximal PSNRs and average PSNR for a whole sequence. The most interesting is the minimal value, which presents the most degraded I frame in a sequence. From the given results, it is possible to see that in most degraded frames a difference in PSNR between proposed method and other two is never bigger then 1.5 dB. The minimal PSNR value of 35.82 dB shows that high fidelity is preserved and that watermarked frames are indistinguishable from originals. The most degraded frame (frame 345 from the Mobile and Calendar sequence) is given with original in Figure 3 for comparison.

The main advantage of the proposed combined method can be seen through a watermark to host ratio (*WHR*) that is watermark signal-to-noise ratio in an embedding window:

$$WHR = N_{nz} \cdot \frac{\mu_\alpha^2}{\overline{x^2}} \qquad (8)$$

where $\overline{x^2}$ is mean AC coefficients power , N_{nz} is number of non-zero AC coefficients, μ_α is mean embedding amplitude.

The most demanding sequence is the "BBC3" sequence. Consisting mainly of low-frequency transitions from black to white and vice-versa, this sequence contains relatively small number of DCT coefficients with high values describing strong edges with high luminescence changes. Considering the low number of TEXTURE blocks (1.76%) and relatively high number of EDGE blocks (35.42%) in the "BBC3" sequence, the Zhang method fails to compete with the other two methods and gives disappointingly small *WHR* (660.70) comparing to Watson (*WHR*=1129.26) and proposed method (*WHR*=1713.43). To

estimate the maximal capacity of the watermarking message, we define signal-to-noise ratio per watermarking bit:

$$\frac{S}{N}[\text{dB}] = 10 \cdot \log_{10} \frac{WHR}{n} \qquad (9)$$

where *n* is the number of embedded bits.

The SNR curves, showing dependence of SNR on the used embedding method and number of bits per embedding window, are given in Figure 4. From the given SNR curves and recalling the SNR boundary of 9.6 dB from section 2, we conclude that in order to achieve bit error rate as low as 10^{-5}, we can embed maximum 72 bits using the Zhang method, 123 bit using Watson method or 188 bits using the proposed combined method for the watermark amplitude adjustment.
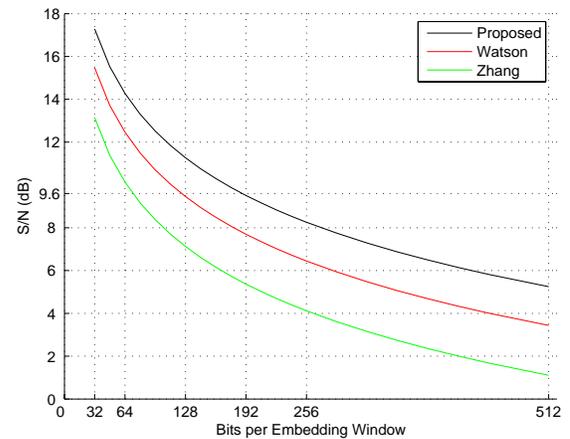

Figure 4. Signal-to-noise Ratio for the first embedding window in the "BBC 3" sequence for three methods

This maximal capacity rates were evaluated by measuring bit error rate. We were embedding 10^{+5} bits with different embedding message sizes (64-256 bits per 8 I frames) in the first embedding window of the "BBC3" sequence using three methods. Messages were extracted and compared with original ones. Measured bit error rates are shown in the Figure 5. It can be seen that measured

values are comparable with estimated ones. We were able to decode 64-bit messages embedded with Zhang method without errors. Using Watson model 128-bit messages were decoded without errors, while as expected the best results are achieved with the proposed method. Here, 192-bit messages were extracted and no errors were observed, which is in consistent with estimated maximal capacity of 188-bits.
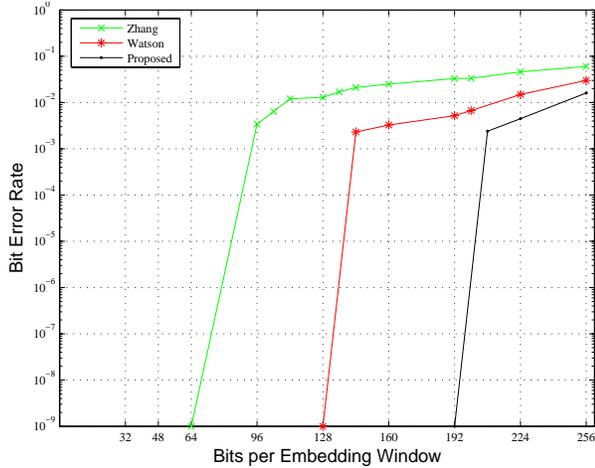

Figure 5. Measured Bit Error Rates for three methods (Z- Zhang, W- Watson, P- proposed)

5. CONCLUSIONS

A comparison of state-of-the-art perceptual models for video watermarking in the DCT domain has been presented. It was shown that Zhang model indeed produces better result than Watson's for textured frames, but achieves considerably less capacity in frames with high number of edges. The new combined method outperforms the two models in terms of watermark capacity while preserving the high imperceptibility of the watermark.

Ongoing and future developments focus on using state-of-the-art turbo codes to increase the watermark capacity. Recent experiments show that it can double the capacity [7].

7. REFERENCES

[1] F. Hartung, and B. Girod, "Watermarking of uncompressed and compressed video," Signal Processing, vol. 66, no. 3, pp. 283-302, 1998.

[2] C. I. Podilchuk and W. Zeng, "Image-adaptive watermarking using visual models," IEEE Journal on Special Areas in Communications, vol. 16, no. 4, pp. 525–539, May 1998.

[3] I. Cox, M. Miller, and J. Bloom, "Digital Watermarking," Morgan Kaufmann Publisher, 2001, 1-55860-714-5.

[4] X.H. Zhang, W.S. Lin and P. Xue, "Imroved Estimation for Just-noticeable Visual Distortions", Signal Processing, vol. 85, no. 4, 2005, pp. 795-808.

[5] L. Cheveau, "Choosing A Watermarking System for Digital Television – The Technology and The Compromises", IBC2002 (www.broadcastpapers.com/asset/IBCEBUWatermarking03. htm)

[6] A.J. Ahumada, H.A. Peterson and A. B. Watson, "An Improved Detection Model for DCT Coefficients Quantization", ", Proc. of SPIE, Human Vision, Visual and Digital Display IV, ed. B. E. Rogowitz, vol. 1913-14, 1993, pp. 191-201.

[7] I. Damnjanovic and E. Izquierdo, "Turbo Coding Protection of Compressed Domain Watermarking Channel", Proc. of IEEE International Conference on Computer as a Tool, Belgrade, Serbia and Montenegro, November 2005.

[8] C. B. Schlegel and L.C. Perez, "Trellis and Turbo Coding", IEEE Press, 2004, 0-471-22755-2.