Journal of Visual Communication and Image Representation 52 (2018) 106–117

Contents lists available at ScienceDirect

Journal of Visual Communication and Image Representation

journal homepage: www.elsevier.com/locate/jvci

A spatiotemporal saliency-modulated JND profile applied to video watermarking

Antonio Cedillo-Hernandez*, Manuel Cedillo-Hernandez, Mariko Nakano Miyatake, Hector Perez Meana

Seccion de Estudios de Posgrado e Investigacion, ESIME Culhuacan, Instituto Politecnico Nacional, Santa Ana Av. 1000, Mexico City 04430, Mexico

ARTICLE INFO

Keywords:
Video watermarking
Visual attention
Saliency regions
Just-Noticeable-Distortion (JND)

ABSTRACT

The just noticeable distortion (JND) has been considered a suitable solution for controlling the watermark strength and generating robust watermarking schemes with distortions that are below the sensitivity threshold. However, JND assumes the same attention level for all image regions, which does not reflect the behavior of an observer. Recently, several models have utilized the modulatory effect of visual attention over JND to improve the efficiency of watermarking schemes. However, most of them have focused on still images. In this paper, we propose a saliency-modulated JND profile for improving video watermarking schemes. Our method aims to adapt the watermark strength to obtain the most robust possible scheme with an imperceptible watermark. Moreover, it has the advantage of fully exploiting the spatiotemporal properties of video to minimize its perceptual redundancies and achieve low computational complexity. Experimental results show the effectiveness of our proposed method and its contributions to video watermarking process.

1. Introduction

DIGITAL watermarking provides a proper platform that aims at protecting copyrighted multimedia data from illegal manipulation and undesired distribution [1,2]. One of the main challenges that are faced by watermarking researchers is to embed a sufficiently robust watermark signal to overcome a wide range of attacks while preserving the visual quality of the host signal [3]. Since the human visual system (HVS) performs the final evaluation of the visual quality of processed multimedia, its properties have been long studied to adjust the watermark strength in terms of visual sensitivity. Some watermarking schemes apply implicit HVS properties such as local contrast [4] and local variance [5,6] to adjust the watermark signal strength while preserving the visual quality. However, they require extra factors to regulate the watermark strength that are empirically determined and vary from one image to another [7]. The just-noticeable distortion (JND) emerges as a suitable solution for this problem, as it represents the highest distortion that is tolerated by the HVS and allows the optimal watermark strength to be set to make variations below the sensitivity threshold [8]. A variety of JND models have evolved over the years, from models in the spatial domain (pixel-wise models) [8–11] to models in the frequency domain (sub-band models) [12–14]. Moreover, the application of JND can be focused on images [14,15] by exploiting only spatial characteristics and videos [16–19] by considering features such as motion compensation and the temporal contrast sensitivity function (CSF).

The main issue with the JND threshold is that it assumes the same attention level on all image areas, which is not consistent with an observer's behavior. Visual attention is an important mechanism that regulates human perception since it identifies those areas that attract the HVS attention (salient regions) and thus modulates the visual periphery [20,21]. Since the observer's capability to detect distortions is higher on salient regions than the remaining areas, the performances of the watermarking schemes can be enhanced if visual attention modulates the JND threshold, i.e., the JND model is adapted based on salient regions. Recently, some works on the application of saliency-modulated JND models to digital watermarking were reported in the scientific literature. Nevertheless, most of them focused on still images.

In this paper, we propose a saliency-modulated JND profile that improves the robustness and imperceptibility of video watermarking techniques. In contrast to current proposals, our method has the advantages of fully exploiting the spatiotemporal properties of HVS and being an agile process. Our full proposal is divided into two main parts: First, we describe our proposed saliency-modulated JND profile, which consists of three steps: (a) JND estimation, (b) saliency mapping, and (c) the modulation stage. In this part, the main contribution of the

[*] Corresponding author.
E-mail addresses: acedillo@itesm.mx (A. Cedillo-Hernandez), mcedilloh@ipn.mx (M. Cedillo-Hernandez), mnakano@ipn.mx (M. Nakano Miyatake), hmperezm@ipn.mx (H. Perez Meana).

https://doi.org/10.1016/j.jvcir.2018.02.007
Received 7 September 2017; Received in revised form 5 January 2018; Accepted 11 February 2018
Available online 16 February 2018
1047-3203/

authors lies not in the detailed parts but in the whole framework that composes the proposed JND profile. The second part of this paper shows how to take advantage of the proposed saliency-modulated JND profile by applying it to video watermarking. Here, we use the proposed JND profile to adjust the energy of a basic video watermarking technique, thereby improving its performance. The performed robustness tests include signal processing and video-based operations. Imperceptibility is measured by using advanced frame- and video-based metrics to obtain a better correlation with HVS subjective quality assessment. Both parts, namely, the design of our saliency-modulated JND model and its application, are compared with some existing solutions to highlight their advantages and contributions. Experimental results confirm the effectiveness of the proposed JND profile, since it is able to compute a JND map in a computationally inexpensive manner and is consistent with observer attention in videos with much and little movement, which allows smart modulation of the watermark strength, thereby producing a more robust watermarked video sequence with high visual quality.

The remainder of this paper is organized as follows: a brief review of related works is presented in Section 2. Section 3 provides a detailed description of the proposed spatiotemporal saliency-modulated profile. The proposed video watermarking scheme, which includes embedding and extraction processes, is presented in Section 4. Section 5 shows and discusses the experimental results, and Section 6 presents the conclusions of this research.

2. Related works

This section presents a brief review of previously published works that are related to the application of saliency-modulated JND models to digital watermarking. Niu et al. [22] proposed a watermarking scheme that is implemented in the discrete cosine transform (DCT) domain, where a saliency-modulated JND profile regulates the watermark strength. The JND profile is calculated by merging contrast and luminance masking, with the spatial CSF and the spectral residual [23]. Wan et al. [24] reported a watermarking technique that is based on the logarithmic spread transform dither modulation framework (STDM). The authors use Watson's model [12] to obtain the JND threshold. Then, the luminance and texture features are merged to compute a saliency map for adjusting the STDM quantization step. Finally, Agarwal et al. [25] employ a saliency-modulated JND profile, in which the JND threshold modulates the scheme energy and a saliency map is used to preserve relevant areas of the image. None of these techniques consider temporal features of the HVS, so they are not suitable for videos. Only a few studies on video watermarking using a saliency-modulated JND profile have been reported. In [26], Cedillo-Hernandez et al. introduced a watermarking scheme for video that uses the quantization index modulation (QIM) model [27]. All frames are divided into regions of interest and regions of no interest by using a visual attention method for still images. Then, the QIM step size is adjusted in each DCT block by using a JND profile only partially. The scheme achieves good robustness against video transcoding operations; however, the method does not fully exploit the temporal properties of HVS. Moreover, the saliency map calculation technique that is used in this reference is highly computationally expensive. Chen et al. [28] proposed a spatiotemporal saliency map for obtaining a location at which to perform the watermark embedding process. This method proposes a fusion of static and temporal saliency maps for exploiting spatiotemporal HVS features, but the static map uses a saliency method that is computationally costly [29,30]. Furthermore, Chen's method is limited to providing a watermark location and does not exploit the features of the full frame.

3. Saliency-modulated JND profile

Perceptual redundancies refer to any visual information that cannot be perceived by the HVS. Thus, their removal does not affect the visual quality of a given signal. In this section, we introduce our proposed saliency-modulated JND profile, which is a process that fully exploits the spatiotemporal HVS properties and represents the perceptual redundancies of video quantitatively. Our proposed saliency-modulated JND profile involves three main steps: JND estimation, saliency mapping, and the modulation stage.

3.1. JND estimation

In this section, we compute the highest error that is tolerated by the HVS for each DCT sub-band. The following explains in detail two JND profiles. Both models are DCT-based, focus on video and have outstanding performance. The next sections describe the performance and contribution of these methods. Jia et al. [18] introduced a JND profile (referred to as the JND-1 profile hereinafter) that is the product of a base threshold, a contrast mask, and a luminance adaptation (1):

$$\mathrm{JND{-}1}(t,n,i,j) = T(t,n,i,j)\cdot\psi_{CONT}(t,n,i,j)\cdot\psi_{LUM}(t,n) \tag{1}$$

where i and j are the DCT subbands indices, n indicates the index of a block in a frame, and t is the frame position along the video sequence. The base threshold T is determined by (2):

$$T(t,n,i,j) = \frac{g}{G(t,n,i,j)\cdot\Delta g}\cdot\eta_{i,j} \tag{2}$$

where G is the spatiotemporal baseline CSF, g is the number of gray levels, and Δg denotes the difference between the maximum and minimum gray intensities on the frame. The term $\eta_{i,j}$ is a compensatory factor that depends on *oblique* and *spatial summation* factors [12], which is defined as:

$$\eta_{i,j} = \frac{1}{\Phi_i\cdot\Phi_j\cdot(r + (1-r)\cdot\cos^2\varphi_{ij})} \tag{3}$$

$$\Phi_u = \begin{cases}\sqrt{1/N}, u = 0\\ \sqrt{2/N}, u > 0\end{cases} u = 0,1,\ldots,N-1 \tag{4}$$

$$\varphi_{ij} = \arcsin\frac{2\omega_{i,0}\omega_{0,j}}{\omega_{i,j}^2} \tag{5}$$

$$\omega_{i,j} = \frac{1}{2N}\cdot\sqrt{(i/\theta_x)^2 + (j/\theta_y)^2} \tag{6}$$

$$\theta_h = 2\cdot\arctan\left(\frac{\Lambda}{2\cdot l}\right), h = x,y. \tag{7}$$

The compensatory factor (3) includes DCT normalization elements, which are denoted as Φ; a parameter r which regulates the *oblique effect* and has a value of 0.6, and the directional angle of a DCT component (φ_{ij}). In this computational model, N is the dimension of the DCT block, $\omega_{\mathrm{i,j}}$ is the spatial frequency of a DCT sub-band, and θ_{x}, and θ_{y} are the horizontal and vertical angles of a pixel respectively. The displayed length of a pixel (Λ) is equal to 1 in most of the displays and the viewing distance l is calculated using the international standard ITU-R BT 500-11 [31]. The baseline CSF measures the HVS acuity as a function of the speed of an object, which is displayed in front of the retina (8):

$$G(t,n,i,j) = c_0(k_1 + k_2|\log(\varepsilon\cdot V(n,t)/3)|^3)\cdot V(n,t)\cdot(2\pi\omega_{i,j})^2\cdot \\ \exp(-2\pi\omega_{i,j}\cdot c_1\cdot(\varepsilon\cdot V(n,t) + 2)/k_3) \tag{8}$$

where $c_0 = 7.12$, $c_1 = 0.56$, $k_1 = 6.1$, $k_2 = 7.3$, $k_3 = 23$, and $\varepsilon = 1.7$ [18]. The retinal velocity V is the difference between the speed of an object within an image when eye movement is not involved (V_{I}) and the eye movement velocity (V_{E}):

$$V(n,t) = V_I(n,t) - V_E(n,t) \tag{9}$$

$$V_E(n,t) = \min[\beta\cdot V_I(n,t) + V_{MIN}, V_{MAX}] \tag{10}$$

$$V_I(n,t) = f \cdot \sqrt{(MV_x(n,t) \cdot \theta_x)^2 + (MV_y(n,t) \cdot \theta_y)^2}. \tag{11}$$

The term $\beta = 0.92$ refers to the efficiency of object tracking. The range values for eye velocity are $V_{MIN} = 0.15$ and $V_{MAX} = 80$ deg/s, and the image velocity V_I is calculated using the video frame rate f and the motion vector components (MV) per block. Contrast masking is computed by merging intra-band (ρ_{Intra}) and inter-band (ρ_{Inter}) masking:

$$\psi_{CONT}(t,n,i,j) = \rho_{Intra}(t,n,i,j) \cdot \rho_{Inter}(t,n) \tag{12}$$

$$\rho_{Intra}(t,n,i,j) = \begin{cases} 1, & (i,j) \in L \cup M \text{for E \& P block} \\ \lambda_1, & \text{otherwise} \end{cases} \tag{13}$$

$$\lambda_1 = \max\left\{1, \left[\frac{C(i,j)}{T(t,n,i,j) \cdot \psi_{LUM}(n,t)}\right]^{\varepsilon}\right\}. \tag{14}$$

DCT blocks are classified into Texture (T), Edge (E) and Plain (P) blocks according to [14]. In this method, DCT sub-bands, which are denoted as $C(i,j)$, are divided in low (L), middle (M) and high (H) frequencies, $\varepsilon = 0.36$, $\lambda_2 = 1 + [((M + H) - \xi_1)/\kappa]\ \lambda_3$, $\lambda_3 = 1.25$, $\lambda_4 = 1.125$, $\xi_1 = 290$ and $\kappa = 1510$:

$$\rho_{Inter}(t,n) = \begin{cases} \lambda_2, & \text{for T blocks} \\ \lambda_3, & \text{for E block,} L + M > 400 \\ \lambda_4, & \text{for E block,} L + M \leqslant 400 \\ 1, & \text{for P block} \end{cases} \tag{15}$$

Finally, the luminance adaptation is computed as follows:

$$\psi_{LUM}(t,n) = \begin{cases} k_4\left(1 - \frac{2 \cdot C(0,0)}{\mu \cdot N}\right)^{\vartheta_1}, C(0,0) \leqslant \frac{\mu \cdot N}{2} \\ k_5\left(\frac{2 \cdot C(0,0)}{\mu \cdot N} - 1\right)^{\vartheta_2}, \text{otherwise} \end{cases} \tag{16}$$

where $k_4 = 2$, $k_5 = 0.8$, $\vartheta_1 = 3$ and $\vartheta_2 = 2$ [18]. Wei et al. [19] propose a spatiotemporal JND model (referred to as the JND-2 profile hereinafter) defined as a product of a spatial threshold and a temporal regulatory factor:

$$\text{JND}-2(t,n,i,j) = T_S(t,n,i,j) \cdot \Omega_T(t,n,i,j). \tag{17}$$

The spatial threshold T_S is given by the product of contrast masking (Ω_C), luminance adaptation (Ω_L), and a basic threshold T_{BASIC} and is defined as:

$$T_S(t,n,i,j) = T_{BASIC}(t,n,i,j) \cdot \Omega_C(t,n,i,j) \cdot \Omega_L(t,n) \tag{18}$$

$$T_{BASIC}(t,n,i,j) = s \cdot \exp(c \cdot \omega_{i,j})/(a + b \cdot \omega_{i,j}) \cdot \eta_{i,j} \tag{19}$$

where $s = 0.25$, $a = 1.33$, $b = 0.11$ and $c = 0.18$ [19]. The spatial frequency ($\omega_{i,j}$) and the compensatory factor ($\eta_{i,j}$) were defined previously in (6) and (3) respectively. Contrast masking is computed by categorizing the DCT blocks into T, E and P blocks using a process that is based on the Canny Operator [32] which is explained in detail in [19]. Then, an elevation agent α for each DCT block type is determined by (20):

$$\alpha = \begin{cases} 1, & \text{for P \& E blocks} \\ 2.25, & \text{for } (i^2 + j^2) \leqslant 16 \text{ in T blocks} \\ 1.25, & \text{for } (i^2 + j^2) > 16 \text{ in T blocks} \end{cases} \tag{20}$$

where i and j are DCT subbands indices. With this value, the contrast masking is calculated using (21):

$$\Omega_C = \begin{cases} \alpha, \text{for} (i^2 + j^2) \leqslant 16 \text{ in P \& E blocks} \\ \alpha \cdot \Psi, \text{otherwise} \end{cases} \tag{21}$$

$$\Psi = \min\left(4, \max\left(1, \left[\frac{C(i,j)}{T_{BASIC}(t,n,i,j) \cdot \Omega_L(t,n)}\right]^{0.36}\right)\right). \tag{22}$$

As the last step in the computation of the spatial threshold, the luminance adaptation effect is given by (23):

Table 1
A comparative between JND-1 and JND-2 models.

JND model	Category (Domain)	Luminance adaptation	Contrast masking	Motion direction	Spatial & temporal CSF
JND-1	Sub-band (DCT)	Yes	Yes	No	Together
JND-2	Sub-band (DCT)	Yes	Yes	Yes	Separately

$$\Omega_L = \begin{cases} (60 - \bar{I})/150 + 1, & \bar{I} \leqslant 60 \\ 1, & 60 < \bar{I} \leqslant 170 \\ (\bar{I} - 170)/425 + 1, & \bar{I} \geqslant 170 \end{cases} \tag{23}$$

where $\bar{I}$ is the average intensity value of the block. Finally, the temporal regulatory factor (Ω_T) that is used in (17) is defined as:

$$\Omega_T = \begin{cases} 1, & \omega_{i,j} < 5\ \&\ \omega_T < 10 \\ 1.07^{(\omega_T - 10)}, & \omega_{i,j} < 5\ \&\ \omega_T \geqslant 10 \\ 1.07^{\omega_T}, & \omega_{i,j} \geqslant 5 \end{cases} \tag{24}$$

where ω_T is the temporal frequency and is given by (25):

$$\omega_T = \frac{i}{2N\theta_x} \cdot V_x + \frac{j}{2N\theta_y} \cdot V_y. \tag{25}$$

The two terms in (25) are a particular case of the estimation of spatial frequency and retinal velocity. If we use $j = 0$ and $\theta_y = 0$ in (6) and (9)-(11) respectively, we obtain the horizontal component (first term), and in contrast, if we use $i = 0$ and $\theta_x = 0$, we obtain the vertical one (second term).

Table 1 presents a comparison of the JND-1 and JND-2 models with the aim of summarizing their similarities and differences. The main similarities are that the two models are DCT-based; they merge luminance adaptation, contrast masking, and a temporal factor; and they can be applied to compute JND for video. On the other hand, we can highlight three main differences between the models: (a) although the JND-1 and JND-2 schemes use similar modulatory factors, such as contrast and luminance, they calculate them in considerably different ways; (b) the JND-2 model considers motion direction, which causes different effects on spatial frequencies; and (c) JND-2 uses a temporal regulatory factor that assumes that the temporal frequency is dependent on the spatial frequency, i.e., it considers temporal and spatial CSF as separable components, which can lead to inaccurate JND estimation [33]. In contrast, in the JND-1 model, both the temporal and spatial CSF effects are formulated as a single multidimensional function.

3.2. Saliency mapping

Visual attention is a complex cognitive process that implies a set of strategies for simplifying the inherent search mechanism in visual perception [34]. Although the research on saliency detection began many years ago, most studies have focused on images and only some computational models for representing the concept of visual attention in video sequences have been constructed [28,30,35–37]. These methods use temporal data from video sequences to obtain a more precise saliency map. However, their process is carried out in the uncompressed domain. Thus, their main drawback is the associated high computational cost. To compute a fast and accurate saliency-modulated JND model, we employ a visual saliency technique that works with spatiotemporal features that are extracted directly from a compressed video stream [38]. The method obtains three features (luminance, color, and texture) from intra-frames and one feature (motion) from inter-frames. First, to compute the static saliency map (S_S), the DC coefficient value in each 8 × 8 DCT block of the luminance (Y) component of each intra-frame represents the Luminance feature (L). In a similar way, color features (C_1, C_2) are computed by using the DC coefficients from Chrominance components (Cr, Cb). According to zig-zag scanning, the first nine AC coefficients of each DCT block of the Y component represent the texture feature (T). Once we obtain the static features, we use a Gaussian model to compute the weight differences among center and surrounding blocks, as follows:

$$S_i^k = \sum_{j \neq i} \frac{F_{i,j}^k}{\sigma\sqrt{2\pi}} \cdot e^{-0.5 \times E_{i,j}^2 \times \sigma^{-2}} \tag{26}$$

where S_i^{k} is the saliency map value in the i_{th} block, that corresponds to the feature k, with $k \in \{L, C_1, C_2, T\}$; $E_{i,j}$ represents the Euclidean distance between the i_{th} and j_{th} DCT blocks; and σ is the standard deviation of the Gaussian distribution. The term $F_{i,j}^k$ is the difference between the i_{th} and j_{th} blocks that corresponds to feature k. This value can be a scalar difference (for color and luminance) or a vector distance (for texture feature). The static saliency map is a combination of the normalized saliency maps of luminance, color, and texture:

$$S_S = \sum_{\forall k} N(S^k)/4 \tag{27}$$

where N is the normalization operation on each saliency map, and S^k represents the complete saliency map using each extracted feature k, $k \in \{L, C_1, C_2, T\}$. The motion saliency map S_M is computed for inter-frames by extracting the MVs from the video bitstream. An MV is denoted by a two-dimensional vector ($v_x(t)$, $v_y(t)$) that is assigned to a block and represents the vertical ($v_x(t)$) and horizontal ($v_y(t)$) motion of the block regarding matched blocks on a reference frame [39]. The motion feature (V) is computed according to the type of inter-frame, by using (28):

$$V = MV_p + (-1) \cdot MV_f \tag{28}$$

For bidirectional predicted frames (B-frames), V utilizes the past reference motion (MV_p) and the future reference motion (MV_f). For predicted frames (P-frames), it uses only the past reference motion and the second term in (28) becomes zero. Considering the above, S_M is computed by applying Eq. (26) with $k \in \{V\}$ and using the Euclidean distance to calculate the difference factor $F_{i,j}^k$. The static and motion saliency maps are merged by using the parametrized normalization, sum and product (PNSP) method to obtain the final saliency map S, (29):

$$S = \beta_1 S_S(t) + \beta_2 S_M(t-1) + \beta_3 S_S(t) S_M(t-1) \tag{29}$$

$$\beta_n = \frac{\sum_{(i,j)} S^k(i,j)}{\sum_{(i,j)} \sqrt{(i-S_{i,e})^2 + (j-S_{j,e})^2} \cdot S^k(i,j)}, n = \{1,2\} \tag{30}$$

$$S_{i,e} = \frac{\sum_{(i,j)} i \cdot S^k(i,j)}{\sum_{(i,j)} S^k(i,j)} \tag{31}$$

$$S_{j,e} = \frac{\sum_{(i,j)} j \cdot S^k(i,j)}{\sum_{(i,j)} S^k(i,j)} \tag{32}$$

where β_1, β_2, and β_3 are parameters for setting the weight of each component, t indicates that the static saliency map is computed using the current intra-frame and $t-1$ indicates that the motion saliency map is computed using the previous inter-frame. The parameter S^k in (30)-(32) is defined as $S^k = \{S_S\}$ to calculate β_1 and as $S^k = \{S_M\}$ to obtain β_2. Then the value of β_3 is given by $(\beta_1 + \beta_2)/2$. The term $S^k(i,j)$ is the saliency map value that corresponds to feature k at coordinate (i,j).

3.3. Modulation stage

Focusing visual attention on a particular location results in a significant reduction of HVS computational resources in other regions [40]. Therefore, the visual attention process modulates the HVS ability to perceive distortions. The visual sensitivity to errors is numerically calculated as the inverse of the JND. Consequently, the JND threshold decreases in salient areas since the visual sensitivity to errors is higher in those regions. Conversely, this value must be increased to allow more distortion in the remaining areas. We use the study that is presented in [41] to generate the modulating function for adjusting the JND profile map in salient areas, which is defined as follows:

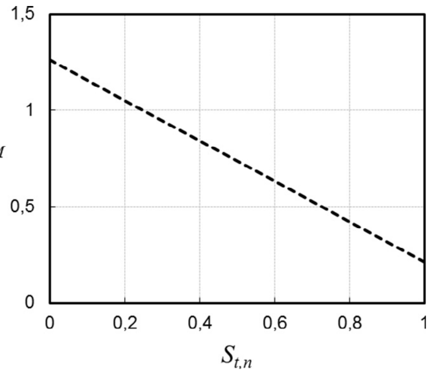

Fig. 1. The modulation function $\Psi^M(t,n)$ which is employed to adjust the JND map regarding the saliency map average in the n_{th} block of the t_{th} frame ($S_{t,n}$).

$$S(\mathrm{JND}) = \mathrm{JND}(t,n,i,j) \cdot \Psi^M(t,n) \tag{33}$$

where S(JND) is the representation of the saliency-modulated JND profile and Ψ^M is an empirical linear function for adjusting the JND value of the n_{th} block of the frame at time t, which is defined as $\Psi^M(t,n) = 1 - (S_{t,n} - \tau_1)\ \tau_2$ and is graphically represented in Fig. 1. The term St,n denotes the saliency data average in the n_{th} block of the t_{th} frame and the parameters τ_1 and τ_2 control the modulation process which, based on our experiments, have values of 0.25 and 1.05 respectively.

4. Video watermarking scheme

In this section, we present a video watermarking scheme, which is guided by our proposed saliency-modulated JND profile. The proposed video watermarking scheme uses DCT coefficients of intra-frames (I-frames) and motion vectors of inter-frames (P and B frames). This information is obtained by performing the video decoding process only partially. The variable length coding (VLC) tables, run-length encoding (RLE) and zig-zag scanning are used to obtain the 64 DCT coefficients of each block from the coded intra-frames. A similar process is implemented to extract the motion vector values for inter-frames [38,42]. Since the proposed method does not require the DCT coefficients of inter-frames or spatial information of intra-frames, it is possible to avoid performing unnecessary decoding operations such as motion compensation or inverse DCT, which are the most computationally expensive operations [43]. In the proposed video watermarking method, two watermark bits are embedded in each 8 × 8 DCT block using only the luminance data of I-frames. To obtain a more robust scheme, the watermark is embedded in the AC sub-bands that are more resilient to the quantization task [26]. The QIM algorithm is the core method for performing the watermark embedding and detection processes. Commonly, the QIM method embeds the watermark data into a host signal by quantizing it with a fixed scalar quantifier. In our proposed approach, this quantifier is dynamically adapted by using the proposed saliency-modulated JND profile. This approach allows a more robust scheme against several attacks to obtained, without significantly affecting the visual quality. Below, we describe the watermark embedding and detection processes.

4.1. Watermark embedding process

Fig. 2 shows the proposed watermark embedding process which is described as follows: (i) Partially decode the video stream to obtain the luminance and chrominance DCT blocks of 8 × 8 pixels from intra-

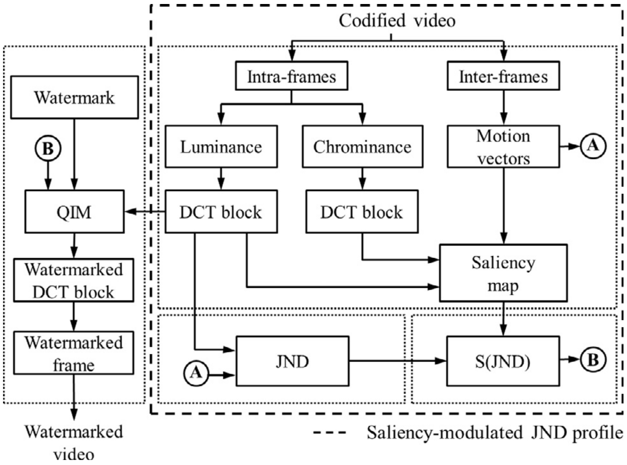

Fig. 2. The proposed watermark embedding process.

frames and the motion vectors from inter-frames. (ii) Obtain the features of luminance, texture, color and motion from extracted video data for computing the saliency map of a frame F at time t, according to (26)-(32). (iii) Calculate the JND thresholds of DCT sub-bands $C_{1,2}$ and $C_{2,1}$ for all DCT blocks of F. (iv) Modulate the previously computed JND threshold of each DCT sub-band by using (33). (v) Build the watermark vector $W = \{w_1, w_2,\dots,w_k\}$, $w_k \in \{0,1\}$, $k \in \{1,2,\dots,K\}$, $K = (M \times N)/32$ where M and N denote the frame dimensions. (vi) Use the QIM algorithm to embed the k_{th} bit of the watermark vector (w_k) in the $C_{i,j}$ sub-band, as follows:

$$C^w_{i,j} = \begin{cases} \mathrm{sgn}(C_{i,j})\cdot\left\lfloor \frac{|C_{i,j}|}{2\alpha_{i,j}} \right\rfloor\cdot 2\alpha_{i,j}, & w_k = 0 \\ \mathrm{sgn}(C_{i,j})\cdot\left(\left\lfloor \frac{|C_{i,j}|}{2\alpha_{i,j}} \right\rfloor\cdot 2\alpha_{i,j} + \alpha_{i,j}\right), & w_k = 1 \end{cases} \tag{34}$$

where $C^w_{i,j}$ represents the $(i,j)_{th}$ watermarked DCT sub-band and $\alpha_{i,j}$ is the QIM quantifier which is defined as:

$$\alpha_{i,j} = S(\mathrm{JND}(t,n,i,j))\cdot Q \tag{35}$$

where Q is the fixed quantization step which is experimentally determined later. The term $S(\mathrm{JND}(t,n,i,j))$ is the saliency-modulated JND threshold for the $(i,j)_{th}$ DCT sub-band of the n_{th} block in frame F at time t. This value is computed in step (iv) of the watermark embedding process. Note that the QIM embedding process (34) is carried out for both the $C_{1,2}$ and $C_{2,1}$ DCT sub-bands.

4.2. Watermark extraction process

Fig. 3 shows a chart of the proposed watermark extraction process which consists of the following steps: (i) The video codified bitstream is partially decoded to extract the 8 × 8 DCT blocks of the luminance space of the watermarked frames. (ii) The watermark vector is obtained by using the QIM extraction process on the $(i,j)_{th}$ watermarked DCT sub-band:

$$\ddot{w}_k = \begin{cases} 0, & \text{if } round(C^w_{i,j}/\alpha_{i,j})=\text{even} \\ 1, & \text{if } round(C^w_{i,j}/\alpha_{i,j})=\text{odd} \end{cases} \tag{36}$$

where $\ddot{w}_k$ is the k_{th} bit of the extracted watermark vector, $C^w_{i,j}$ is the $(i,j)_{th}$ watermarked DCT sub-band and $\alpha_{i,j}$ is the QIM quantifier, which is computed in the embedding process and stored to be used as a secret key in the extraction process. (iii) The bit error rate (BER) is used to determine the reliability of the extracted watermark with respect to the embedded one.

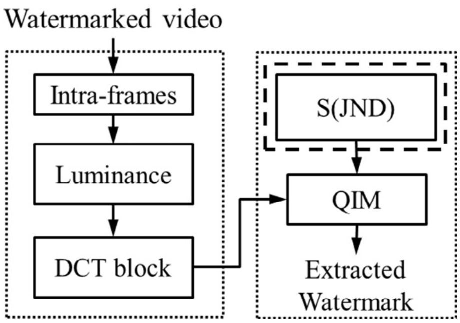

Fig. 3. The proposed watermark extraction process.

5. Experimental results

In this section, we carry out experiments to evaluate the effectiveness of our proposed saliency-modulated JND profile and to measure its contribution to enhancing the watermarking process. To conduct our tests, we have built a video database of twenty videos, which are codified on the MPEG-4 Part 2 compression standard on Advanced Simple Profile [44]. The first ten videos have a resolution of 352 × 288 pixels (CIF format) at 30 frames per second (fps). The rest of the videos have a size of 704 × 576 pixels (4CIF format) at 24 fps. The test database is composed of low- and high-motion sequences with different conditions of lighting, texture, and color. All results were obtained from a prototype that was implemented using the MATLAB© platform.

5.1. Evaluation of the saliency-modulated JND profile

We evaluate the performance of our proposed JND profile from three perspectives: (1) time-consumption, (2) consistency of the obtained saliency map with the observer's attention in environments with much and little movement, and (3) how much noise the method is able to embed into a video without being detected by the HVS.

Time-consumption analysis: One of the most important drawbacks of current approaches is the large amount of computational time that is required to compute the saliency-modulated JND profile. In this section, we conduct a performance comparison between existing works and our proposed approach in terms of processing time. For fair comparison, we consider three stages for all methods: video decoding process, JND computing, and saliency estimation. Table 2 shows the average amount of time that is required to process all video sequences at each stage. The results are divided in terms of video size to distinguish processing time of CIF and 4CIF formats. The total processing times for all videos are shown in the last column of Table 2. According to Table 2, the video decompression time that is required by our proposed approach is approximately 10% of the times for the methods in [26,28]. This notable difference is due to the partial decoding process that is used in our proposed approach. All methods employ very similar JND techniques. The main difference is that the method in [26] performs the JND calculation for all the frames of the video sequence, while the method in [28] and our proposed approach perform the JND calculation only in key frames. Finally, the visual attention based on information maximization (AIM) [45] method, which is used in [26],

Table 2
Time consumption analysis.

Method	Video Size	Time (minutes)				
		Decoding	JND	Saliency	Total	Average
[26]	4CIF	6.1	9.7	15.6	31.4	21.3
	CIF	3.1	3.4	4.7	11.3	
[28]	4CIF	6.1	1.7	3.7	11.5	8.1
	CIF	3.1	0.6	1.1	4.8	
Proposed	4CIF	0.5	1.3	1.4	3.2	2.2
	CIF	0.3	0.5	0.4	1.1	

has a high computational cost, as it can take up to 15 min to process all frames of a video with 4CIF spatial resolution. Although the method in [28] and our proposed approach perform saliency estimation only for key frames, the saliency estimation method that is employed in our proposed approach [39] reduces the computation time by approximately 40% compared to the model of visual attention to salient proto-objects [46], which is used in [28]. The final average shows that our proposed JND profile achieves significant time savings since it is approximately three times faster than the method in [28] and nine times faster than the model in [26].

Saliency evaluation: The effectiveness of a saliency-modulated JND profile strongly depends on its capacity for calculating a saliency map that is consistent with the factors that guide attention. The more accurate the calculation of salient areas, the higher the chance to take advantage of unsupervised regions and obtain a more efficient watermarking scheme. We test the performance of our JND profile by using high- and low-motion videos. For this purpose, we classify all the video sequences according to the level of movement activity. From encoded video data, we can obtain the number of macroblocks that are classified as Motion Compensated (MC) or Non-Motion Compensated (No_MC) within an inter-frame (P-frame). The presence of a No_MC macroblock in a P-frame indicates a perfect coincidence with its reference (skipped block or not coded at all) or a large difference that cannot be expressed with motion estimation and has to be inter-coded (scene changing) [47]. The absence of No_MC blocks indicates the presence of motion within a video scene. Therefore, since the video motion intensity is inversely proportional to the number of No_MC macroblocks in a P-frame, we define the video motion ratio μ as:

$$\mu = \frac{\text{Number of inter } \mathit{NoMC} \text{ Macroblocks}}{\text{Number of Frame Macroblocks}}. \tag{37}$$

In this way, a higher value of video motion ratio μ indicates a video sequence with little movement. By using this metric, we classify all videos as high- or low-motion videos according to whether the ratio is below or above the database average. With this classification, we analyze the saliency map that is obtained in each case by using two popular metrics: (1) the Normalized Scanpath Saliency (NSS) [48] and (2) the Area Under the ROC Curve (AUC) [49]. We evaluate the results with more than one metric to guarantee that the conclusions are independent of the chosen metric. The reader is referred to [50] to obtain a detailed explanation of these saliency evaluation metrics. The binary fixation map (a zero matrix with 1 s at locations of fixations), which is used to compute both saliency metrics, is obtained according to the classification of each video. For videos with little motion, we consider static elements and perform semi-automatic separation of foreground objects from the background. In contrast, for high-motion videos, the binary fixation map is automatically obtained by considering the blocks with larger motion vectors. Then, we calculate the values of the NSS and AUC metrics among the binary fixation map and the computed saliency maps ([26,28], and our proposed approach) by using MATLAB functions, which are freely available in [51]. The obtained results are shown in Table 3.

According to Table 3, the AIM method, which is used in [26], achieves the best performance for static saliency maps (videos with little motion) according to both the NSS and AUC metrics. A value of NSS that is closer to one indicates larger saliency values at human binary fixated locations than other locations. Similarly, the larger the AUC score, the better the saliency prediction of the saliency map. For motion saliency maps (videos with high motion), our proposed approach obtains the best results for NSS and AUC, which means that our proposed approach can better determine areas with greater motion activity. The model of visual attention to salient proto-objects, which is used in [28], does not achieve good performance since the computed salient areas are insufficient to cover the binary fixation map. An average of all the computed saliency maps, including static and motion maps, is shown in the last column of Table 3. As our scheme achieves the best average scores for the NSS and AUC metrics, we can assert that our proposed method estimates a saliency map that is consistent with the features that guide attention in high- and low-motion videos. This characteristic is one of the key aspects of the proposed approach for enhancing the robustness and imperceptibility of video watermarking methods.

Table 3
Saliency maps precision evaluation.

Method	Metric	Saliency Map		Average
		Static	Motion	
[26]	NSS	0.92	0.72	0.82
	AUC	0.85	0.62	0.74
[28]	NSS	0.49	0.41	0.45
	AUC	0.54	0.50	0.52
Proposed	NSS	0.87	0.93	0.90
	AUC	0.80	0.81	0.81

To better illustrate the achieved performance, we show the results for one representative sequence of high- and low-motion videos (Fig. 4). The upper row of Fig. 4 corresponds to the “soccer” video sequence, which has a motion ratio of $\mu = 10.5\%$, i.e., it is a high-motion video. As we stated before, the binary fixation map (second column) is built by considering the blocks with larger motion vectors. In this case, the AIM method, which is used in [26], (third column) achieves scores of 0.79 and 0.60 for the NSS and AUC metrics, respectively. The visual saliency model, which is used in [28], (fourth column) obtains values of 0.61 and 0.50 for the NSS and AUC metrics, respectively, since there are very few salient areas. The last column shows the saliency estimation results of our proposed approach, with values of 0.95 and 0.87 for the NSS and AUC metrics, respectively. The saliency map that is computed by our proposed approach matches almost perfectly the binary fixation map. Finally, the lower row of Fig. 4 shows the “container” video sequence, which has a motion ratio of $\mu = 96\%$. Here, the best performance is achieved by the AIM method (NSS = 0.96, AUC = 0.88), followed by the proposed scheme (NSS = 0.87, AUC = 0.80) and the model of visual attention to salient proto-objects (NSS = 0.43, AUC = 0.5). Although our proposed approach does not achieve the best performance for static saliency maps, it obtains suitable representations of salient regions, which are mostly coincident with the static binary fixation map.

Noise injection capability: The performance of a JND profile is evaluated in terms of its ability to generate a noise-injected frame with similar visual quality on a higher level of noise. Thus, the method must better distribute the noise across the frame. In the third experiment, we create a noise-injected video by using the JND models that are described in tion 3.1. Then we modulate the noise over it by applying our proposed JND model. The aim of this test is to determine which method generates the lowest visual perceptual distortion. The model for producing a noise-injected video is as follows:

$$C'(t,n,i,j) = C(t,n,i,j) + R(n,i,j)\cdot f_{\text{JND}} \tag{38}$$

where C' is the noise-injected DCT sub-band on the n_{th} block of the frame at time *t*. The term *R* takes the values of $(+1, -1)$ in a random way and f_{JND} denotes the JND threshold. The four JND estimators that are applied to obtain the JND threshold f_{JND} are denoted as (a) JND-1, (b) *S*(JND-1), (c) JND-2, and (d) *S*(JND-2). To obtain a convincing evaluation, the visual quality is judged in two ways: (i) by evaluating the global quality distortion and (ii) using a block-based distortion measure to confirm that the noise is injected in appropriate regions of the frame. The global distortion is measured by using the PSNR metric. Table 4 shows the obtained results after computing the PSNR between original and noise-injected videos. The presented values correspond to the average of the frames of all video sequences.

According to Table 4, the JND-1 and JND-2 profiles are improved by

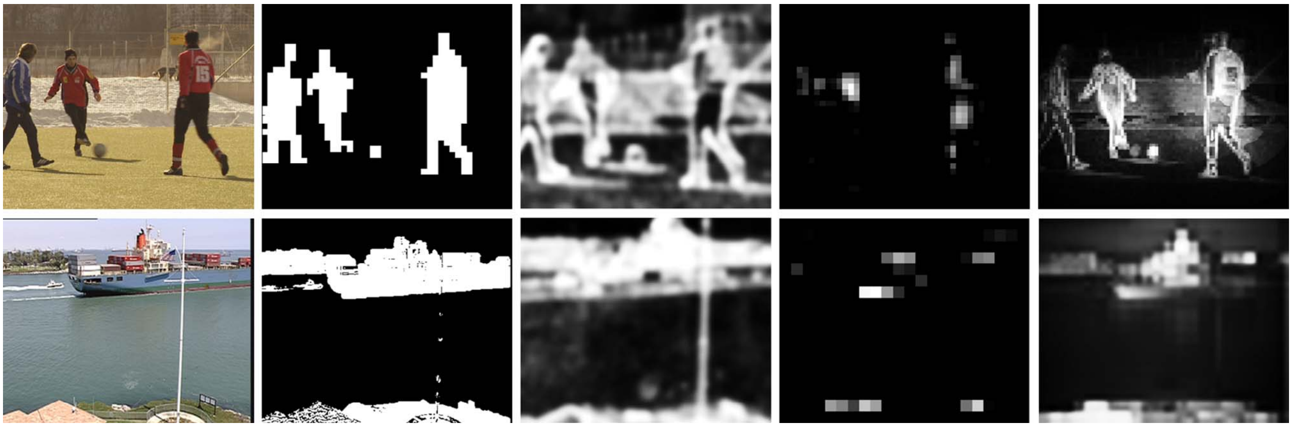

Fig. 4. An evaluation of the precision of saliency maps. From left to right, original frame from video sequence; binary fixation map; saliency map from AIM model [45] used in [26], saliency map from the visual attention model to salient proto-objects [46] used in [28], and the saliency map used in our proposal [39].

Table 4
Visual quality achieved by different JND profiles.

	f_{JND}			
	JND-1	*S*(JND-1)	JND-2	*S*(JND-2)
PSNR (dB)	36.18	34.09	33.84	29.85

our saliency-modulated JND model, since the video frames can accept more distortion (an average gain of approximately 3 dB) without it being easily perceived by a viewer. The proposed method modulates the injected noise by exploiting the HVS boundaries with a strategic approach, i.e., it achieves a more aggressive distortion on the less notable regions of each frame. To prove the above, we first examine the injected noise in the most relevant areas. Fig. 5 shows a magnified view of the 30th frame of the "Miss-America" video sequence. The magnified view matches the highest-sensitivity area, i.e., the area with the highest values in the saliency map. In the bottom-left corner of each image in Fig. 5, a magnified version of the area that is bounded by a red[1] square is shown. The modulatory effect of our proposed approach (Fig. 5(c) and (e)) decreases the distortion on the most important parts of the frame, thereby improving the perception of visual quality. Moreover, it is noticeable that the JND-1 model (Fig. 5b) has a better performance than JND-2 (Fig. 5d) since it produces a noise-injected frame with less-perceptible distortion.

To determine the distribution of the noise across the entire image, we carry out an assessment of the quality distortion by using the 5th frame of the "flower" video sequence (Fig. 6(a)). Each frame is divided into blocks of 32 × 32 pixels. Then, the structural similarity index (SSIM) [52] between the original and the noise-injected blocks is calculated. The SSIM metric has better correlation with the subjective evaluation of the quality of the HVS than the PSNR. Fig. 6(c) is the noise-injected frame that is generated by using the JND-1 model and Fig. 6(e) is its modulated version, which is produced by using our proposed JND profile. For SSIM maps (Fig. 6(d) and (f)), brighter blocks have higher SSIM values (better visual quality). According to Fig. 6, the JND-1 scheme achieves good modulation of noise (Fig. 6(d)) since it uses several properties of the HVS, such as block classification, contrast masking, and luminance adaptation, as we explained in Section 3. However, our proposed approach achieves better distribution of noise by reflecting higher values of SSIM on the observer's attention area (Fig. 6(b)) and a greater amount of noise in the remaining regions (Fig. 6(f)).

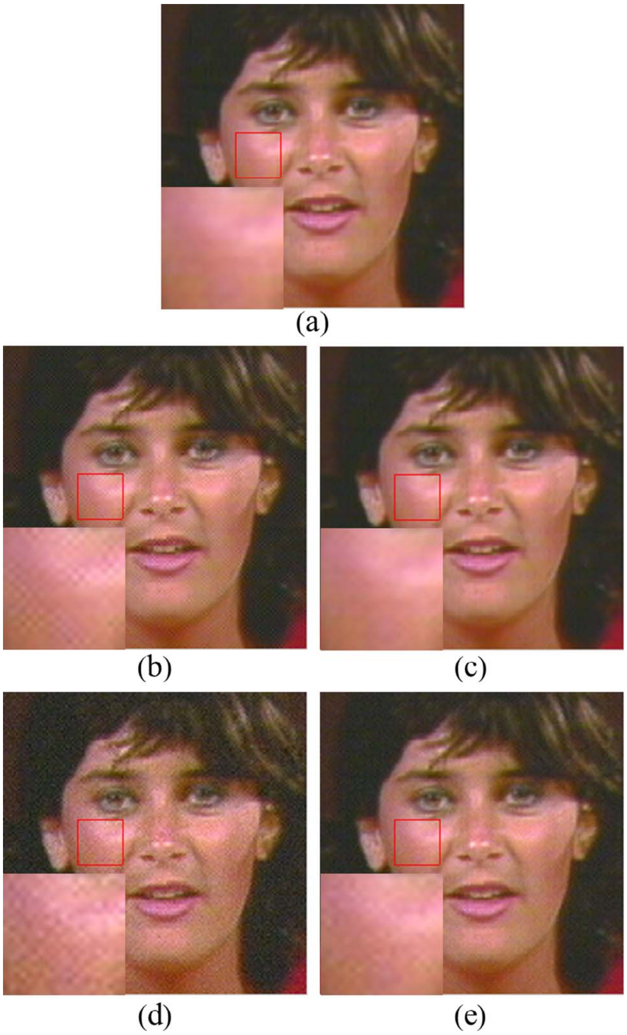

Fig. 5. The most sensitive area of the 30th frame of the "Miss-America" video sequence. (a) Original frame, (b) Noise-injected frame using JND-1 with PSNR = 34.85 dB, (c) JND-1 modulated by our proposed approach (S(JND-1)) with PSNR = 32.09, (d) Noise-injected frame using JND-2 with PSNR = 30.24, and (e) JND-2 modulated by our approach (S(JND-2)) with PSNR = 28.12.

5.2. Evaluation of video watermarking framework

In this section, we assess the contributions of our proposed saliency-modulated JND profile when it is applied in the video watermarking

[1] For interpretation of color in Fig. 5, the reader is referred to the web version of this article.

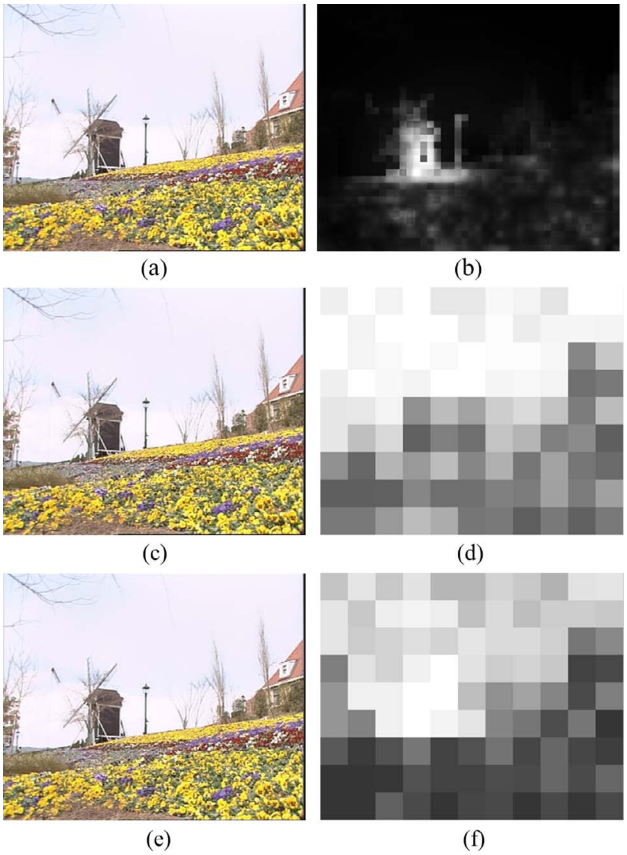

Fig. 6. A block-based quality distortion evaluation. (a) 5th original frame of the “Flower” video sequence, (b) Saliency map of (a), (c) Noise-injected frame using JND-1, (d) SSIM map of (c), (e) JND-1 noise-injected frame modulated by our proposed approach, and (f) SSIM map of (e).

process. All the experiments of this section are based on the strategy of performing the watermark embedding process (Section 4.1) on all videos by varying the watermark energy, i.e., the QIM quantifier ($\alpha_{i,j}$), which is defined in (35).

For this purpose, we define a “basic watermarking model” (Model I) as a method in which the QIM quantifier is a fixed value, i.e., all the DCT sub-bands are watermarked with the same strength. Then, the Model I is modified by modulating the watermark strength with the JND-1 (Model II) and the JND-2 (Model IV) profiles. The Models II and IV are referred to as “unmodulated JND profiles” since they use a JND profile to better distribute the watermark energy, but do not consider the visual attention. Finally, Models II and IV are enhanced by using the proposed saliency-modulated JND profile to distribute the watermark energy while taking into account salient regions. Model III corresponds to the saliency-modulated version of Model II and Model V is the modulated version of Model IV. Table 5 presents a summary of the QIM quantifier values as well as the Model ID for each test set. Subsequently, we conduct several experiments to evaluate the achieved performances on each test set in terms of robustness and imperceptibility.

(1) Parameter determination: Here, we set the value of the QIM fixed quantifier (Q) which allows a proper initial watermarking strength to be obtained for Model I. This value is empirically chosen based on the trade-off between imperceptibility and robustness. To obtain an optimal value for Q, we perform the watermark embedding process, described in Section 4.1, on all video sequences using $\alpha_{i,j} = Q$, with $Q \in \{2\ldots20\}$. Then, we calculate the SSIM and the PSNR between the original and watermarked frames to determine how much the visual quality was affected. In addition, we estimate the robustness in terms of BER value between the extracted and embedded watermarks under the same conditions. Table 6 presents the obtained results from this experiment, which correspond to the averages over all frames of all video sequences. From Table 6, we have selected $Q = 12$ as the most appropriate value according to the trade-off between robustness and imperceptibility (boldface data line). With this initial strength value, PSNR and SSIM values of 50.02 dB and 0.9901, respectively, are achieved, which indicate that the watermark is imperceptible to an observer since, by definition, the value of SSIM is 1 when two images are identical. A BER value of zero indicates that the watermark has been entirely recovered under these conditions.

Table 5
QIM quantifier values to conduct comparisons.

Model ID	QIM quantifier ($\alpha_{i,j}$)
Model I	Q
Model II	$Q \times$ JND-1(t,n,i,j)
Model III	$Q \times S$(JND-1(t,n,i,j))
Model IV	$Q \times$ JND-2(t,n,i,j)
Model V	$Q \times S$(JND-2(t,n,i,j))

Table 6
Determination of QIM fixed quantifier for Model I.

Q	PSNR (dB)	SSIM	BER
2	55.42	0.9989	0.0058
4	54.89	0.9980	1.01 E−05
6	54.22	0.9967	1.01 E−05
8	52.84	0.9950	7.36 E−07
10	51.41	0.9926	4.20 E−09
12	**50.02**	**0.9901**	**0.0000**
14	48.79	0.9899	0.0000
16	47.67	0.9867	0.0000
18	46.68	0.9832	0.0000
20	45.75	0.9792	0.0000

(2) Watermark imperceptibility: To determine the quality of watermarked videos, the PSNR is obtained between original and watermarked frames (Fig. 7). The PSNR value for Model I is 50.02 dB, which was defined in the previous section. Models II and IV have values of 39.03 dB and 37.94 dB, respectively. These values are higher than those that are shown in Table 6 since only two DCT sub-bands are affected by the watermark embedding process. The PSNR values of frames that are modulated by our proposed approach, in Model III and Model V, are 36.43 dB and 34.32 dB, respectively. Here, we apply a similar judgment to that used in Section 5.1 where, although a lower PSNR value is obtained, the difference is viewed as a gain since our proposed approach improves the quality of critical areas of the frame and allows greater distortion in less significant regions. Our proposed saliency-modulated JND profile (Models III and V) produces an average gain of 14.6 dB compared to the basic watermarking method (Model I) and an average

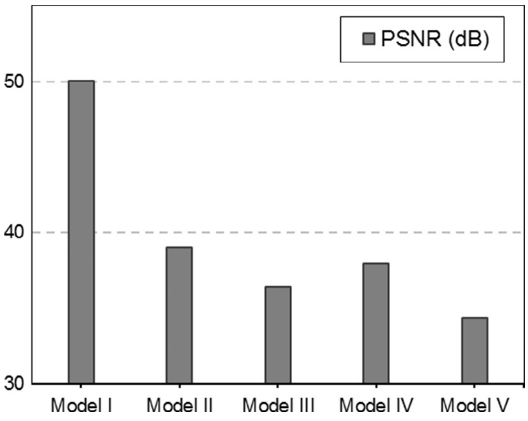

Fig. 7. The PSNR average values between original and watermarked videos.

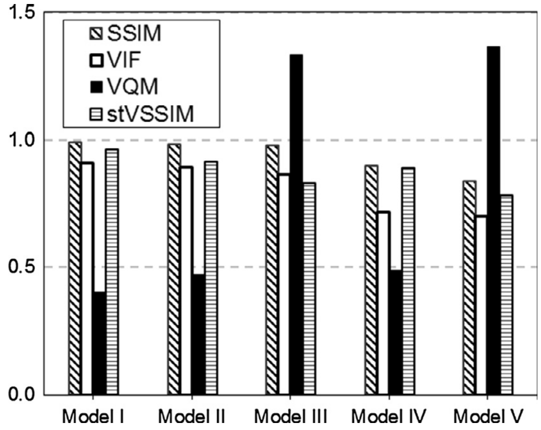

Fig. 8. The average obtained results after applying the SSIM, VIF, VQM, and stVSSIM quality metrics between original and watermarked videos.

gain of 3.11 dB compared to the unmodulated JND profiles (Models II and IV).

However, it is well known that the correlation between the PSNR and the HVS evaluation of quality is not sufficiently close. For this reason, we measure the visual quality of watermarked videos by using four HVS-based quality metrics that achieve an approximation that is closer to human perception. Fig. 8 shows the results that are obtained by comparing original and watermarked videos with the following metrics: the SSIM, Visual Information Fidelity (VIF) [53], the Video Quality Model (VQM) [54], and the spatiotemporal video SSIM (stVSSIM) [55]. The SSIM and VIF metrics are frame-based and their values decrease as the visual quality is affected.

According to Fig. 8, the differences between the basic watermark model (Model I) and the other evaluated models (Models II–V) are significantly lower with advanced quality metrics than those that are obtained with the PSNR. These results suggest that the quality distortion is barely noticeable and reaffirms the idea of gain that is described at the end of the previous paragraph. From Fig. 8, Model II and its saliency-modulated version (Model III) achieve high-quality performances for both the SSIM and VIF metrics (Model II: SSIM = 0.98, and VIF = 0.89; Model III: SSIM = 0.97, and VIF = 0.86). These scores suggest that the watermark is not perceptible by a viewer. In contrast, Models IV and V obtain smaller values of SSIM and VIF (Model IV: SSIM = 0.89, VIF = 0.71; Model V: SSIM = 0.83, VIF = 0.69). Thus, they produce a higher distortion of visual quality.

Frame-based metrics are suitable for measuring the quality of low-motion videos; however, for high-motion videos, it is relevant to consider the inherent temporal redundancy by using video-based quality metrics. VQM and stVSSIM are video-based metrics that consider temporal distortions.

A higher value of stVSSIM indicates good quality, while a smaller value of VQM represents a greater fidelity to the original video. According to Fig. 8, the scores of the VQM and stVSSIM metrics support the results that were obtained by SSIM and VIF, which means that an observer cannot easily perceive the quality distortion of videos that contain high- and low-motion content. The VQM values are 0.4, 0.47, 0.48, 1.33, and 1.36 for Models I to V, while the stVSSIM values are 0.96, 0.91, 0.82, 0.89, and 0.78 for Models I to V, respectively.

To better illustrate the distortion that is caused by the watermark embedding process, in Fig. 9, we show the original 150th frame of the "Silent" video (Fig. 9(a)), its saliency map (Fig. 9(b)), and its watermarked versions that are obtained using Model I (Fig. 9(c)), Model II (Fig. 9(e)), and Model III (Fig. 9(g)). In addition, we show the embedded watermarks in Fig. 9(d), (f) and (h), respectively. Although the watermark is embedded in the DCT domain, we display it in the spatial domain with sharpness enhancement for visual convenience. According to Fig. 9, the watermark that is embedded by Model I (Fig. 9(d)) is barely noticeable. The Model II works properly by raising the watermark energy in textured regions and reducing it in plain areas (Fig. 9(f)). However, the best performance is achieved by the proposed JND profile (Fig. 9(h)), as it reduces the watermark strength in salient areas and enhances it in the remaining regions.

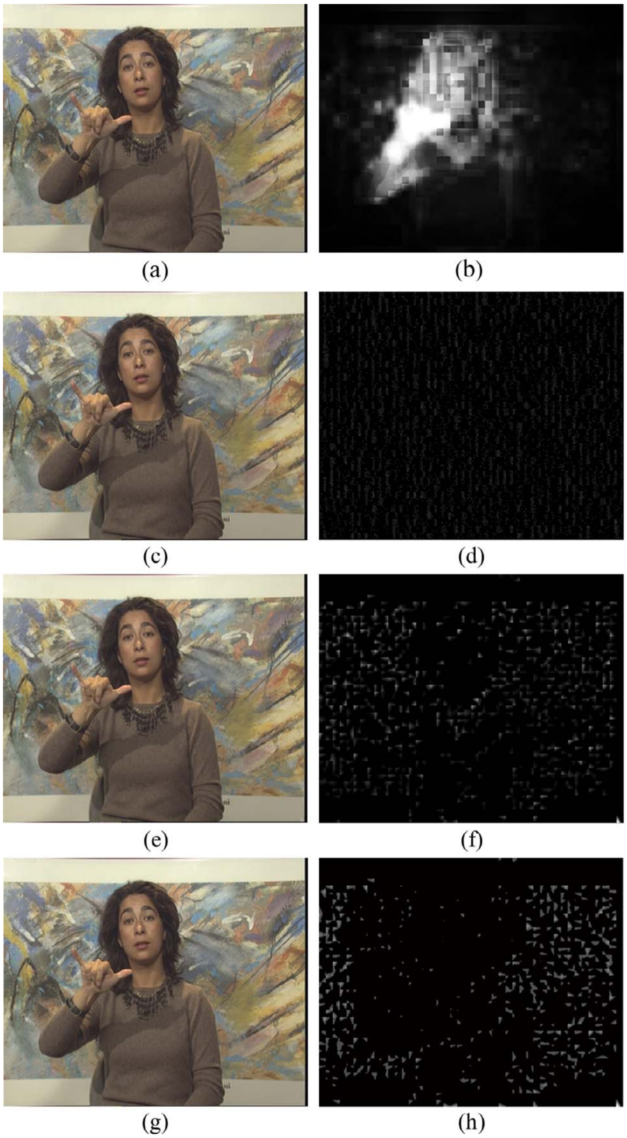

Fig. 9. (a) The original 150th frame of "Silent" video, (b) saliency map of (a), its watermarked versions using (c) Model I, (e) Model II, and (g) Model III, and the embedded watermarks of (d) Model I, (f) Model II and (h) Model III.

(3) Watermark Robustness: Under practical circumstances, a watermarked video sequence can be manipulated by carrying out hostile intentional or nonintentional tasks. Such processes can partially or completely remove the embedded watermark from a video sequence. The proposed watermarking framework was designed to embed the watermark data in the most resilient DCT subbands, which provides high resistance to various attacks. We assess the watermark robustness against common signal processing and video-based operations. We simulate all the hostile tasks by incrementing the level of distortion on watermarked videos. Then, we analyze the distorted video to compute the BER and determine the robustness in each case. Fig. 10 shows the performances of Models I-V against signal processing attacks. The performance results that are achieved against impulsive and Gaussian noise contamination are shown in Fig. 10(a) and (b), respectively. The density of the impulsive noise varies from 0 to 0.010 and the Gaussian noise variance ranges from 0 to 0.008. These attacks can eventually

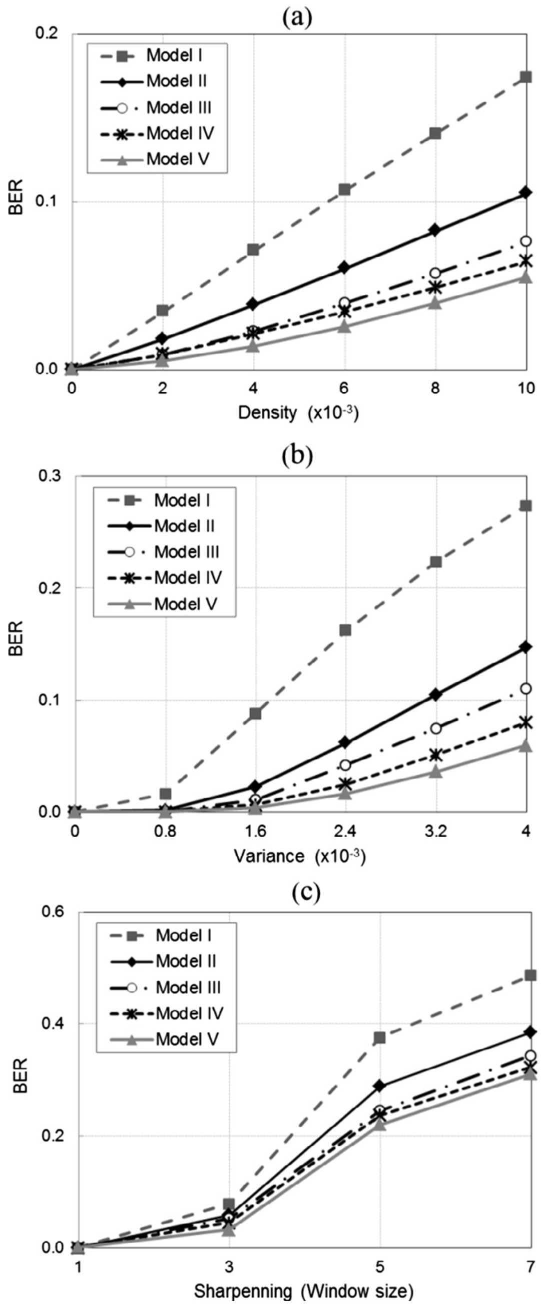

Fig. 10. The watermark robustness performance against (a) impulsive noise contamination, (b) Gaussian noise contamination, and (c) sharpening.

Table 7
Robustness against video compression standard and bit rate changing.

Compression standard	Bit rate	BER				
		Model I	Model II	Model III	Model IV	Model V
H.264-AVC	4 Mbps	0.061	0.015	0.008	0.006	0.003
	2 Mbps	0.125	0.029	0.010	0.009	0.006
	1 Mbps	0.203	0.064	0.023	0.019	0.017
MPEG-2	4 Mbps	0.061	0.013	0.005	0.004	0.000
	2 Mbps	0.161	0.022	0.007	0.007	0.005
	1 Mbps	0.224	0.056	0.012	0.011	0.009
WMV	4 Mbps	0.059	0.015	0.008	0.006	0.001
	2 Mbps	0.166	0.023	0.009	0.008	0.008
	1 Mbps	0.221	0.056	0.012	0.011	0.010
VP6	4 Mbps	0.072	0.012	0.005	0.002	0.000
	2 Mbps	0.174	0.025	0.008	0.007	0.005
	1 Mbps	0.227	0.059	0.014	0.011	0.011

cause significant visual distortion. For example, values of PSNR = 32.5 dB and SSIM = 0.63 are reached for an impulsive noise attack with a density of 0.004. As expected, Model I obtains the highest values of BER. Thus, it is the weakest method among those tested. For both attacks, the Models II and IV show significantly enhanced performance compared to Model I since the JND-1 and JND-2 profiles take advantage of the HVS properties to embed a stronger watermark. Furthermore, the improvement that is achieved by our proposed JND profile when it is applied to modulate Models II and IV is noteworthy. Model III and Model V perform better not only by improving the imperceptibility of the watermark, as we demonstrate in the previous section but also by using a more robust scheme against noise contamination attacks. Fig. 10(c) shows the results of robustness against sharpness attacks, which are frequently performed in an intentional way. The window size of the sharpness operation varies from 1 to 7. A sharpness operation with a window size that is larger than 3 is not usually performed since it produces high distortion of visual quality. However, we apply these values to obtain a wider comparative range. In contrast to noise contamination, sharpness attacks result in smaller differences in performance among Models II to V. Nevertheless, the models that are modulated by our proposed approach achieve lower BER values.

A common operation that is performed by end-users is to change the features of a video to adapt it to different devices. The two main features that an end-user wants to change are (a) the video format, since not all devices can reproduce all formats, and (b) the large size of video files, which is widely related to bit-rate change. These operations are aggressive and pose challenges to the preservation of the watermark information in video sequences. For this test, the original videos are codified by using the MPEG-4 Part2 compression standard with an average bit rate of 8 Mbps. These videos are watermarked by using Models I-V. Then, to evaluate the robustness against video-based attacks, we use a transcoder to change the compression standard and the bitrate. Table 7 shows the results of this operation, including the compression standard, the bit rates, and the calculated BER values in each case. According to Table 7, Model I does not achieve good robustness performance since it obtains high BER values. Model II enhances Model I and achieves very low BER values in all cases. The quantization process is utilized to decrease the bit rate and, as a consequence, the visual quality of videos. Model III obtains lower BER values than Model II and comparable values to those that are obtained by Models IV and V. Here, the small values of BER make accurate watermark extraction possible after compression standard and bit rate changes. Considering the complete results of Table 7, we conclude that Model III achieves the best performance for video-based hostile operations in terms of imperceptibility and robustness.

(4) Performance comparison: With the aim of highlighting the relevance of the obtained results, in this section, we perform a performance comparison among Models III and V of our proposed approach and the results that are obtained by the methods that are proposed in [26,56]. For fair comparison, we selected recently proposed schemes that focus on video watermarking and whose experiments are performed under similar conditions to those that are used to evaluate our proposed approach (e.g., similar test video database and simulations attacks). Furthermore, these methods use the same metrics to measure imperceptibility and robustness. Table 8 shows the BER results after assessing the watermarked videos by each method against three representative attacks: (a) MPEG compression with a bit-rate of 1 Mbps, (b) H.264 compression with a bit-rate of 1 Mbps and (c) impulsive noise contamination (Noise I) with a density of 0.02. According to Table 8, Model V achieves the best performance against video MPEG conversion, as it obtains the lowest value of BER (0.009). For this attack, Model III and the scheme in [26] also achieve small values of BER that signify a retrieval error of approximately 1% of the watermarking information.

Table 8
Robustness performance comparison.

	No attack			BER	
	PSNR	SSIM	MPEG	H.264	Noise I.
Model III	36.43	0.97	0.012	0.023	0.017
Model V	34.32	0.83	0.009	0.017	0.005
[26]	33.66	0.81	0.014	0.024	0.019
[56]	36.88	0.87	0.055	0.000	0.055

Similar performances are achieved by all methods against the impulsive noise contamination operation. Again, Model V obtains the best performance and Model III and [26] achieve BER values that are close to 0.02. In the two previous attacks, the scheme in [56] obtains the highest values of BER (0.055), which indicates the worst performance. In contrast, for H.264 video compression, the method that is proposed in [56] achieves the best performance since the watermark is recovered in its entirely. However, Model V obtains a low BER value of 0.017, and Model III and the scheme in [26] only fail to retrieve approximately 2% of the watermark data. Overall, the efficacy of a watermarking method must be measured by considering the trade-off between robustness and imperceptibility. The left side of Table 8 shows the imperceptibility results (in terms of PSNR and SSIM) of the watermarked frames in the absence of attacks. The method in [26] obtains values of PSNR = 33.66 and SSIM = 0.81. This method uses a strategy to obtain a robust scheme against aggressive transcoding operations and it considers some HVS characteristics. However, temporal HVS properties are not taken into account in their entirely and the method is computationally expensive. In contrast, the main objective of the approach in [56] is to achieve real-time performance by reducing the computational cost. To accomplish this, it employs a window that is localized at the center of the frame as the target for watermark embedding. This method obtains values of PSNR = 36.88 and SSIM = 0.87 and, despite obtaining higher values than the method in [26], the strategy does not guarantee the imperceptibility of the watermark in video frames with large plain areas. Method V of our proposed approach achieves PSNR = 34.32 and SSIM = 0.83. Although it is the most robust method, it does not obtain the highest level of imperceptibility. This may be due to some inherent features of the JND-2 model, which we described at the end of Section 3.1. Finally, as we mentioned in the previous section, the best relationship between imperceptibility and robustness is achieved by Model III, which uses our proposed JND profile. Model III obtains PSNR = 36.43 and SSIM = 0.97, and considering that SSIM quantifies similarity with human perception, the obtained value (SSIM = 0.97) indicates the highest visual quality among those evaluated, with good performance in terms of robustness.

6. Conclusions

We have presented the performance results of the proposed saliency-modulated JND profile and measured its contribution to improving the video watermarking process. Our proposed approach provides a fast and accurate method to fully exploit the spatiotemporal HVS properties and create noise-injected video frames of similar visual quality under a higher level of noise. We investigate the modulatory behavior of the visual attention process over current video-oriented JND profiles to improve the watermark trade-off between imperceptibility and robustness. The above has been experimentally demonstrated by measuring the visual distortion that is caused by our proposed approach with frame-based and video-based metrics, which confirm the transparency of the watermark. Furthermore, several experiments are carried out to assess the robustness of our proposed video watermarking framework against signal processing and video-based hostile operations. Our experimental results indicate that our proposed approach produces a more robust video watermarking scheme with an average gain of 14.6 dB compared with basic watermarking methods and 3.11 dB compared with unmodulated JND profiles. Thus, our proposed approach can be considered a suitable option and a potential research direction for improving video watermarking schemes.

Acknowledgment

The authors would like to thank the Instituto Politecnico Nacional for support in the realization of this research.

References

[1] R.L. Lagendijk, G.C. Langelaar, I. Setyawan, Watermarking digital image and video data, IEEE Signal Process. Mag. 17 (5) (2000) 20–46.
[2] Y. Tew, K. Wong, An overview of information hiding in H. 264/AVC compressed video, IEEE Trans. Circuits Syst. Video Technol. 24 (2) (2014) 305–319.
[3] S.A. Parah, J.A. Sheikh, N.A. Loan, G.M. Bhat, Robust and blind watermarking technique in DCT domain using inter-block coefficient differencing, Digital Signal Process. 53 (1) (2016) 11–24.
[4] Q.B. Do, A. Beghdadi, M. Luong, P.B. Nguyen, A perceptual pyramidal watermarking technique, in: Proc. IEEE ICME, Hannover, Germany, June 2008, pp. 281–284.
[5] F. Bartolini, M. Barni, V. Cappellini, A. Piva, Mask building for perceptually hiding frequency embedded watermarks, in: Proc. IEEE ICIP, Chicago, IL, USA, October 1998, pp. 450–454.
[6] M. Cedillo-Hernandez et al., Security enhancement of medical imaging via imperceptible and robust watermarking, IEICE Transactions on Inf. And Syst., 2015, vol. E98-D, no. 9, pp. 1702–1705.
[7] P.B. Nguyen, A. Beghdadi, M. Luong, Perceptual watermarking using pyramidal JND maps, in: Proc. of 10th IEEE International Symposium on Multimedia, Berkeley, CA, USA, 2008, pp. 418–423.
[8] C.H. Chou, Y.C. Li, A perceptually tuned subband image coder based on the measure of just-noticeable-distortion profile, IEEE Trans. Circuits Syst. Video Technol. 5 (6) (1995) 467–476.
[9] Y.J. Chin, T. Berger, A software-only video codec using pixel-wise conditional differential replenishment and perceptual enhancements, IEEE Trans. Circuits Syst. Video Technol. 9 (3) (1999) 438–450.
[10] X. Yang, W. Lin, Z. Lu, E. Ong, S. Yao, Motion-compensated residue preprocessing in video coding based on just-noticeable-distortion profile, IEEE Trans. Circuits Syst. Video Technol. 15 (6) (2005) 742–750.
[11] C.H. Chou, C.W. Chen, A perceptually optimized 3-D subband codec for video communication over wireless channels, IEEE Trans. Circuits Syst. Video Technol. 6 (2) (1996) 143–156.
[12] B. Watson, DCTune: A Technique for visual optimization of DCT quantization matrices for individual images, in: Soc. Inf. Display Dig. Tech. Papers XXIV, 1993, pp. 946–949.
[13] A.J. Ahumada, Jr., H.A. Peterson, Luminance-model-based DCT quantization for color image compression, in: Proc. SPIE Human Vision, Visual Processing, and Digital Display III, vol. 1666, B. E. Rogowitz, Ed., 1992, pp. 365–374.
[14] X. Zhang, W.S. Lin, P. Xue, Improved estimation for just-noticeable visual distortion, Signal Process. 85 (4) (2005) 795–808.
[15] A.B. Watson, G.Y. Yang, J.A. Solomon, J. Villasenor, Visibility of wavelet quantization noise, IEEE Trans. Image Process. 6 (8) (1997) 1164–1175.
[16] D.H. Kelly, Motion and vision. II. Stabilized spatio-temporal threshold surface, J. Opt. Soc. Amer. 69 (1979) 1340–1349.
[17] S. Daly, Engineering observations from spatiolvelocity and spatiotem-poral visual models, Proc. SPIE 3299 (1998) 180–191.
[18] Y. Jia, W. Lin, A.A. Kassim, Estimating just-noticeable distortion for video, IEEE Trans. Circuits Syst. Video Technol. 16 (7) (2006) 820–829.
[19] X. Wei, K.N. Ngam, Spatio-temporal just noticeable distortion profile for grey scale image/video in DCT domain, IEEE Trans. Circuits Syst. Video Technol., vol. 19, no. 3, pp. 337–346, March 2009.
[20] S.J. Luck, M.A. Ford, On the role of selective attention in visual perception, Proc. Nat. Acad. Sci. 95 (3) (1998) 825–830.
[21] Z.K. Lu, W. Lin, X.K. Yang, E.P. Ong, S.S. Yao, Modeling visual attention's modulatory aftereffects on visual sensitivity and quality evaluation, IEEE Trans. Image Process. 14 (11) (2005) 1928–1942.
[22] Y. Niu, M. Kyan, L. Ma, A. Beghdadi, S. Krishnan, Visual saliency modulatory effect on just noticeable distortion profile and its application in image watermarking, Signal Process. Image Commun. 28 (8) (2013) 917–928.
[23] X. Hou, L. Zhang, Saliency detection: a spectral residual approach, Proc. IEEE Conf. Computer Vision Pattern Recognition, 2007.
[24] W. Wan, J. Liu, J. Sun, C. Ge, X. Nie, Logarithmic STDM watermarking using visual saliency-based JND model, Electron. Lett. 51(10) (2015) 758–760.
[25] H. Agarwal, D. Sen, B. Raman, M. Kankanhalli, Visible Watermarking based on importance and just noticeable distortion of image regions, Multimedia Tools and Applications, pp. 1–25, 2015.
[26] A. Cedillo-Hernandez, M. Cedillo-Hernandez, M. Garcia-Vazquez, M. Nakano Miyatake, H. Perez-Meana, A. Ramirez Acosta, Transcoding resilient video watermarking scheme based on spatio-temporal HVS and DCT, Signal Process. 97 (2014) 40–54.
[27] B. Chen, G.W. Wornell, Digital watermarking and information em-bedding using

dither modulation, in: Proc. IEEE Workshop Multimedia Signal Process., Redondo Beach, CA, December 1998, pp. 273–278.

[28] D. Chen, S. Xia, K. Lu, A JND-based saliency map fusion method for digital video watermarking, in: Proc. of 34th IEEE control conference Chinese (CCC), Hangzhou, China, July 2015, pp. 4568–4573.

[29] D. Walther, C. Koch, Modeling attention to salient proto-objects, Neural Netw. 19 (2006) 1395–1407.

[30] C. Guo, L. Zhang, A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression, IEEE Trans. Image Process 19 (1) (2010) 185–198.

[31] ITU, Methodology for the subjective assessment of the quality of tele-vision pictures, Geneva, Switzerland, ITU-R BT.500-11, 2002.

[32] J. Canny, A computational approach to edge detection, IEEE Trans. Pattern Anal. Mach. Intell., vol. PAMI-8, no. 6, pp. 679–698, December 1986.

[33] N. Bandekar, A perceptually tuned model for applications to scalable video coding, Ph.D. dissertation, Indian Institute of Technology, Bombay, 2009.

[34] J.K. Tsotsos, Motion understanding: task-directed attention and representations that like perception with action, Int. J. Comput. Vis. 45 (3) (2001) 265–280.

[35] Y. Zhai, M. Shah, Visual attention detection in video sequences using spatio-temporal cues, in: Proc. ACM Int. Conf. Multimedia, 2006, pp. 815–824.

[36] L. Itti, P. Baldi, Bayesian surprise attracts human attention, Adv. Neural Inform. Process. Syst. 46 (8–9) (2006) 1194–1209.

[37] X. Hou, L. Zhang, Dynamic visual attention: searching for coding length increments, Adv. Neural Inform. Process. Syst., vol. 21, pp. 681–688, MIT Press, 2008.

[38] Y. Fang, W. Lin, Z. Chen, C. Tsai, C. Lin, A video saliency detection model in compressed domain, IEEE Trans. Circuits Syst. Video Technol. 24 (1) (2014) 27–38.

[39] L. Duan, T. Xi, S. Cui, H. Qi, A.C. Bovik, A spatiotemporal weighted dissimilarity-based method for video saliency detection, Signal Process. Image Commun. 38 (2015) 45–56.

[40] S. Treue, J.C.M. Trujillo, Feature-based attention influences motion processing gain in macaque visual cortex, Nature 399 (1999) 575–579.

[41] L. Itti, J. Braun, C. Koch, Modeling the modulatory effect of attention on human spatial vision, in: T.G. Dietterich, S. Becker, Z. Ghahramani (Eds.), Advances Neural Information Processing Systems, MIT Press, Cambridge, MA, 2002, vol. 14.

[42] ISO/IEC JTC1, Information technology—Coding of audio-visual objects Part2: Visual, ISO/IEC 14492–2, (MPEG-4 Visual), Version 1: April 1999, Version 2: February 2000, Version 3: May 2004.

[43] F. Cavalli, R. Cucchiara, M. Piccardi, A. Prati, Performance analysis of MPEG-4 decoder and encoder, in: Proc. Int. Symp. Video/Image Process. Multimedia Commun., 2002, pp. 227–231.

[44] (2016, Sept.) Test Video Sequences Web Page [Online]. Available: < https://sites.google.com/site/researchvideosequences/ > .

[45] N. Bruce, J. Tsotsos, Saliency, attention and visual search: an information theoretic approach, J. Vis., vol. 9, no. 3, 2009.

[46] D. Walther, C. Koch, Modeling attention to salient proto-objects, Neural Netw. 19 (9) (2006) 1395–1407.

[47] M. Ghanbari, Standard codecs: image compression to advanced video coding, The Institution of Electrical Engineers, London, U.K., 2003.

[48] R. Peters, A. Iyer, L. Itti, C. Koch, Components of bottom-up gaze allocation in natural images, Vision Res. 45 (18) (2005) 2397–2416.

[49] T. Judd, F. Durand, and A. Torralba, A benchmark of computational models of saliency to predict human fixations, Massachusetts Inst. Technol., Cambridge, MA, USA, Tech. Report, 2012.

[50] A. Borji, H.R., Tavakoli, D.N., Sihite, L. Itti, Analysis of scores, datasets, and models in visual saliency prediction, ICCV, Sidney, Australia, December 2013, pp. 921–928.

[51] (6, Sept.) Saliency Evaluation Measures Webpage [Online]. Available: < https://sites.google.com/site/saliencyevaluation/ > .

[52] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: from error measurement to structural similarity, IEEE Trans Image Process 13 (4) (2004) 600–612.

[53] H.R. Sheikh, A.C. Bovik, Image information and visual quality, IEEE Trans. Image Process. 15 (2) (2006) 430–444.

[54] M.H. Pinson, S. Wolf, A new standardized method for objectively measuring video quality, IEEE Trans. Broadcast. 50 (3) (2004) 312–322.

[55] A K. Moorthy, A.C. Bovik, Efficient motion weighted spatiotemporal video SSIM index, in: Proceedings of SPIE, 7527(1), 2010, pp. 75271I–75271I-9.

[56] I. Bayoudh, S.B. Jabra, R. Zagrouba, A robust video watermarking for real-time application, in: International Conference on Advanced Concepts for Intelligent Vision Systems, Springer, Cham, 2017, pp. 493–504.