VIDEO WATERMARKING USING QUANTIZATION INDEX MODULATION IN THE 3D-DCT DOMAIN

Patrizio Campisi and Alessandro Neri

Dip. Elettronica Applicata
Università degli Studi di Roma TRE
Via della Vasca Navale 84, 00146 Roma, Italy
Email: (campisi, neri)@uniroma3.it

ABSTRACT

A video watermarking system using quantization index modulation in the three-dimensional discrete cosine transform (3D-DCT) domain is here presented. After having partitioned a video sequence into spatio-temporal units, they are projected onto the 3D-DCT domain. A set of transformed coefficients are taken according to a selection rule, ordered according to a predefined scan order, and marked using Quantization Index Modulation watermarking (QIM). A perceptual criterion is also used to define the “local” quantization steps to be used in QIM. Finally, the inverse 3D inverse DCT (IDCT) is performed thus obtaining the marked video shot. Robustness of the proposed method against a variety of both intentional and unintentional video attacks has been observed through experimental results.

1. INTRODUCTION

In this paper we deal with the problem of copyright protection of digital video sequences by using digital watermarking. This issue is rising more and more concerns because of the development of the digital multimedia market and of the always decreasing price of storage devices and digital supports. Extensive surveys dealing with video watermarking applications, attacks and methodologies can be found in [1], [2].

The video temporal dimension is not exploited in the first developed video watermarking methods, where a video is considered as a succession of digital still images. Therefore existing schemes employed for still image watermarking have been employed by watermarking each frame independently of the others. A real time watermarking algorithm robust to video coding, D/A and A/D conversion, format conversion and limited editing, using a bidimensional spread spectrum approach operating in the spatial domain is presented in [3]. A spread spectrum approach to embed digital watermarks in both the uncompressed domain and the compressed MPEG-2 domain is presented in [4]. An attempt to exploit the static and dynamic temporal components of a video has been performed in [5] where a wavelet transform is applied along the temporal axis of the video to design a robust watermarking system. In order to make the embedding of the mark imperceptible, methods exploiting perceptual information have been presented in the recent past (see for example [6], [7]). In [7] a perceptual watermarking method, operating in the 2D-DCT domain, for embedding the mark into MPEG-2 compressed video streams, has been presented. However, in order to design video watermarking methods robust to both non-hostile and hostile video processing operations such as video format conversion, temporal desynchronization, video editing, and video collusion frame-by-frame watermarking strategy are not appropriate. In [8] collusion issues are discussed when frame-by-frame embedding strategies are enforced for video watermarking. In the recent past, to overcome the drawbacks related to frame-by-frame watermarking methods, approaches operating in three dimensional transform domains have been recently proposed. A watermarking approach operating in the 3D discrete Fourier transform domain has been proposed in [9]. A watermarking method operating in the wavelet domain and making use of BCH codes and 3D interliving is presented in [10]. In [11] a watermarking method operating in the 3D-DWT domain, by using pseudo random linear statistics of pseudo random connected regions in the DC subband of the 3D wavelet domain is described. In [12] a perceptual based video watermarking technique, operating in the 3D-DWT domain, that jointly exploits both the spatial and the temporal dimension of a video sequence is described. Quantization based methods are used in [13] to mark 3D subbands in the 3D-DWT domain. In [14] a video watermarking technique using the 3D discrete cosine transform has been described.

Attempts to account for human visual attention in order to better localize video’s regions where to transparently embed the mark have been recently used to design saliency-based watermarking techniques. In [15] the payload is embedded in those regions that receive less attention by a
viewer, so that the perception of the scene is not altered. Among the factors impacting the human visual attention there are: contrast, color, object size and shape, motion, and many more. Although several factors have been identified their relative importance is not yet determined. However, in many studies, the temporal features are considered as characterizing the importance of a region in terms of human perception of quality [16]. For instance, foreground moving objects are prone to attract viewers’ attention making impairments in the background more difficult to detect. At the same time, when objects move faster with respect to the average motion of the scene, impairments in the object would have minor impact on the subject quality assessment.

This would allow to increase the watermark strength in “low”-motion and “high”-motion areas, while reducing it for regions with “medium” temporal activity.

In this paper we propose a novel video watermarking technique operating in the 3D-DCT domain, thus jointly exploiting both the spatial and the temporal dimension of a video sequence. Specifically, the video is projected onto the 3D-DCT domain, and the coefficients selection is performed by taking into account the energy compaction property of the DCT transform. Eventually, the selected coefficients set is marked by using quantization index modulation (QIM) watermarking [17]. This approach, which is gaining always increasing consensus with respect to other approaches such that spread spectrum based watermarking methods, consists in quantizing the selected coefficients by means of a quantizer chosen in a set on the basis of the coefficient to embed. We propose to perceptually design the quantization steps by using a qualitative Human Visual System (HVS) model. After a brief review of the 3D-DCT and of its inverse transform given in Section 2, the proposed watermarking method is detailed in Section 3. Experimental results and conclusions are drawn in Sections 4 and 5 respectively.

2. 3D-DCT

Given a 3D real valued function \( f[n_1, n_2, n_3] \), being \( n_1, n_2 \), and \( n_3 \) integers assuming values in the set \( 0 \leq n_1 < N_1, 0 \leq n_2 < N_2 \), and \( 0 \leq n_3 < N_3 \), its forward 3D-DCT [19], \( F[k_1, k_2, k_3] \), is given by:

\[
F[k_1, k_2, k_3] = \alpha(k_1, k_2, k_3) \sum_{n_1=0}^{N_1-1} \sum_{n_2=0}^{N_2-1} \sum_{n_3=0}^{N_3-1} f[n_1, n_2, n_3] \cdot \cos \left( \frac{2n_1 + 1}{2N_1} k_1 \pi \right) \cos \left( \frac{2n_2 + 1}{2N_2} k_2 \pi \right) \cos \left( \frac{2n_3 + 1}{2N_3} k_3 \pi \right)
\]  

being

\[
\alpha(k_1, k_2, k_3) = \sqrt{\frac{8}{N_1 N_2 N_3}} C(k_1) C(k_2) C(k_3)
\]  

and

\[
C(k_i) = \begin{cases} 
\frac{1}{\sqrt{2}} & k_i = 0 \\
1 & \text{otherwise}
\end{cases}
\]

with \( 0 \leq k_i < N_i \), being \( i = 1, 2, 3 \).

The 3D IDCT [19], is given by:

\[
f[n_1, n_2, n_3] = \sum_{k_1=0}^{N_1-1} \sum_{k_2=0}^{N_2-1} \sum_{k_3=0}^{N_3-1} \alpha(k_1, k_2, k_3) F[k_1, k_2, k_3] \cdot \cos \left( \frac{2n_1 + 1}{2N_1} k_1 \pi \right) \cos \left( \frac{2n_2 + 1}{2N_2} k_2 \pi \right) \cos \left( \frac{2n_3 + 1}{2N_3} k_3 \pi \right)
\]

where \( \alpha(k_1, k_2, k_3) \) and \( C(k_i) \) are given by (2) and (3) respectively.

It is worth pointing out that the 3D-DCT (3D-IDCT) is a separable transform, which implies that it can be implemented as a series of 1D-DCT (1D-IDCT). Therefore its computation can be performed by means of the fast algorithms which have been devised for computing the 1D-DCT (1D-IDCT). The 3D-DCT (3D-IDCT) computational complexity and computing time can be further reduced when fast 3D-DCT (3D-IDCT) algorithms are employed as discussed in [20].

3. THE PROPOSED WATERMARKING METHOD

The video sequence is first segmented into non-overlapping spatio-temporal plot units, called shots, that depict different actions. Each shot is characterized by no significant changes in its content which is determined by the background and the objects present in the scene. Although several video segmentation algorithms which detect the abrupt scene changes [22], [23], [24], have been proposed in literature, for the sake of simplicity, we have segmented the video sequence into shots of fixed length. Then each shot is projected onto the 3D-DCT domain and the coefficients where to embed the mark are selected and ordered according to a specific criterion. The selected coefficients are marked using QIM, and eventually the 3D-IDCT is performed, thus obtaining the marked shot. Specifically, let us consider the \( m \)-th video sequence shot, \( f_m[n_1, n_2, n_3] \), with \( 0 \leq n_1 < N_1, 0 \leq n_2 < N_2 \), and \( 0 \leq n_3 < N_3 \), composed by \( N_3 \) frames of dimension \( N_1 \times N_2 \) pixels. Let us indicate with \( F_m[k_1, k_2, k_3] \) the 3D-DCT of \( f_m[n_1, n_2, n_3] \) obtained using (1).

3.1. Host coefficients selection

The host coefficients where to embed the mark must be chosen in such a way to guarantee both the robustness and the imperceptibility of the mark. Therefore, the low frequency
coefficients, which usually present the higher energy, are left unaltered to guarantee transparency. To achieve robustness, similar rule applies to the high frequency coefficients, which are usually drastically quantized by video encoders due to their low energetic content. Following this criterion, the watermark is embedded in the DCT coefficients that correspond to the middle spatial and temporal frequencies. Specifically, the coefficients choice is performed by selecting those coefficients, in the 3D-DCT volume, whose indices \((k_1, k_2, k_3)\) belong to the region \(R\) limited by the two hyperboloids \((k_1+1)(k_2+1)(k_3+1) = C_l\) and \((k_1+1)(k_2+1)(k_3+1) = C_u\), being \(C_l < C_u\) two properly chosen constants.

### 3.2. Perceptual QIM watermark embedding

The mark embedding is based on the use of the QIM approach [17], which relies on the quantization of the selected host coefficients by means of a quantizer chosen among a set of quantizers according to the data to be embedded. However, to minimize the distortion perception produced by the watermarking, the quantizers can be designed taking into account the spatio-temporal properties of the HVS. Specifically, the HVS has low contrast sensitivity for high horizontal \((k_1)\), vertical \((k_2)\), and temporal \((k_3)\) frequencies, and even lower sensitivity for high joint frequencies \((k_1-k_2, k_1-k_3, k_2-k_3, k_3-k_2-k_3)\). Therefore the quantization step should increase with an increase of \((k_1), (k_2)\), and \((k_3)\), and increase even more for \((k_1-k_2, k_1-k_3, k_2-k_3, k_3-k_2-k_3)\) [18]. These qualitative considerations on the HVS allow us properly choosing the quantization step as a function of the coefficient index \((k_1, k_2, k_3)\) as follows:

\[
\Delta(k) = \Delta(k_1, k_2, k_3) = \left\lfloor \frac{\Delta}{4}(1 + k_1^2 + k_2^2 + k_3^2) \right\rfloor. \tag{5}
\]

Once the embedding region has been selected, its coefficients are ordered in a vector \(\mathbf{V}_R = [v_1, v_2, \ldots, v_L]\), according to a predefined scan order, which we assume known at the receiving side. In our approach the coefficients scan is made by picking the coefficients, belonging to the region \(R\), first along the direction \(k_1\), then along the direction \(k_2\), and finally along the direction \(k_3\). This scan order is determined by the observation that in most cases the coefficients are spread over the smaller values of \(k_3\). Correspondingly, the vector \(\Delta_R = [\Delta_1, \Delta_2, \ldots, \Delta_L]\), whose elements are defined through (5), defines the proper quantization steps to be used for the selected coefficients.

Given a binary representation \(b = \{b_i, 1 \leq i \leq L\}\) of the watermark, and the host coefficients \(\mathbf{V}_R = [v_1, 1 \leq i \leq L]\) the perceptual QIM system we define consists of quantizing \(v_i\) by letting \(b_i\) choose between one of two uniform quantizers \(U_0\) and \(U_1\) defined as:

\[
U_0 = \{u = l\Delta_i + d | l \in \mathbb{Z}\} \quad U_1 = \{u = l\Delta_i + \Delta_i/2 + d | l \in \mathbb{Z}\}\tag{6}
\]

being \(\Delta_i \in \Delta_R\) the quantization step and \(d \in [0, 1)\) a key used to improve security. It is straightforward to note that this method can be expressed in terms of additive watermarking. In fact, by posing

\[
q_i = Q_{\Delta_i} \left[ v_i - \Delta_i \left( \frac{b_i}{2} + d \right) \right] - \left[ v_i - \Delta_i \left( \frac{b_i}{2} + d \right) \right], \tag{7}
\]

being \(Q_{\Delta_i}\) a uniform quantizer of step \(\Delta_i\), the watermarked data can be expressed as

\[
v_i^w = v_i + q_i. \tag{8}
\]

The so obtained marked coefficients, are first grouped in the vector \(\mathbf{V}_R^w = [v_1^w, v_2^w, \ldots, v_L^w]\), then inserted back in the selected region \(R\) in the DCT volume according to the defined scan order, thus obtaining the marked 3D-DCT volume \(F_m[k_1, k_2, k_3]\). Eventually, the 3D-IDCT is performed according to (4) thus obtaining the marked \(m\)-th video sequence shot \(f_m[n_1, n_2, n_3]\).

### 3.3. Watermark extraction

At the receiving side, the \(m\)-th shot of the received marked video sequence, \(f_m[n_1, n_2, n_3]\), is 3D-DCT transformed. The region \(R\), where the mark has been embedded, is then selected according to the aforementioned criterion. The coefficients belonging to the region \(R\) are then extracted and ordered in the vector \(\hat{\mathbf{V}}_R^w = [\hat{v}_1^w, \hat{v}_2^w, \ldots, \hat{v}_L^w]\). Given the generic element \(\hat{v}_i^w\) belonging to \(\hat{\mathbf{V}}_R^w\) and the quantization steps vector \(\Delta_R = [\Delta_1, \Delta_2, \ldots, \Delta_L]\), the decoded value \(\hat{b}_i\) is obtained by first extracting \(y_i\) according to the following rule:

\[
y_i = Q_{\Delta_i} \left[ \hat{v}_i^w - d\Delta_i \right] - \left[ \hat{v}_i^w - d\Delta_i \right], \tag{9}
\]

and then by using hard-decision decoding. Because of the adopted embedding rule (7), the values \(y_i\), should be close to zero when the bit \(b_i\) to be embedded is equal to 0 and close to \(\pm \Delta_i/2\) when \(b_i = 1\). Therefore, the decoded value \(\hat{b}_i\) can be obtained according to the following decision rule:

\[
\hat{b}_i = \begin{cases} 0 & |y_i| < \Delta_i/4 \\ 1 & |y_i| \geq \Delta_i/4. \end{cases} \tag{10}
\]

### 4. EXPERIMENTAL RESULTS

Experimentations have been performed on uncompressed videos composed by \(N_3 = 384\) frames each of size \((N_1 \times \)
The employed test sequences comprise news programs, computer generated videos, musical videos, as well as sport video clips. Each video has been partitioned into shots having fixed length equal to 20 frames. The region $R$ limited by the two hyperboloids, where to embed the watermark, is specified by considering $C_l = 300$ and $C_u = 400$, thus selecting a middle-frequency region in the 3D-DCT domain. A watermark composed by 600 bits, drawn from a uniform distribution, is embedded in the selected region by using the approach described in Section 3. Specifically, the embedding is performed by means of (7)-(8) using a value for $\Delta$ in (5) equal to 185. The values of both the constants $C_l, C_u$ and the parameter $\Delta$ have been empirically chosen, on a wide range of videos, in order to trade off between the imperceptibility of the mark and its robustness against most intentional attacks.

The imperceptibility of the mark has been tested by using the Video Quality Metric (VQM) Software [25] used to measure the perceptual effects on the video of impairment such as blurring, block distortion, unnatural motion, noise and error blocks. Specifically, the general model metric, denoted as $VQM_G$, which can assume values comprised in the interval $[0, 1]$, has been used. Higher values of the metric correspond to higher distorted scenes. For the values of the parameters $C_l, C_u, \Delta$ used in the experimentations a value $VQM_G = 0.2$ has been observed (see Table 1). This value guarantees the imperceptibility of the mark as can be verified by direct observation of the marked video sequence. From Table 1 it is evident how the video quality degrades when both the embedding region became closer to the origin of the 3D-DCT domain, thus involving lower frequency coefficients, and when the quantization step increases.

The robustness of the proposed method has been tested using a variety of unintentional and intentional video attacks such as MPEG-2 and MPEG-4 compression, collusion, transcoding, and frame dropping. As a figure of merit, the ratio $\rho$ between the number of correctly decoded watermark bits and their total number is computed. For the sake of comparison, the method described in [14], operating in the 3D-DCT domain, has been implemented and the experimental results have been here reported. In Figs.1, 2 the performance of both the proposed method and the one presented in [14] in terms of the ratio $\rho$ versus the bit rate of the marked video sequence compressed using both MPEG-4 and MPEG-2 coder have been reported. From Figs.1, 2 we observe that our approach is more robust both against MPEG-2 and MPEG-4 compression than the one proposed in [14]. Type I and type II collusion video attacks have also been considered. As for the first attack a subsequence obtained by picking one frame belonging to the marked video sequence every ten is generated. Then, an estimation of the mark is obtained by averaging the subsequence’s frames. Type II collusion is performed by averaging two consecutive frames in a subsequence of ten frames.

<table>
<thead>
<tr>
<th>$C_l$</th>
<th>$C_u$</th>
<th>$\Delta$</th>
<th>$VQM_G$</th>
</tr>
</thead>
<tbody>
<tr>
<td>200</td>
<td>300</td>
<td>285</td>
<td>0.30</td>
</tr>
<tr>
<td>300</td>
<td>400</td>
<td>185</td>
<td>0.20</td>
</tr>
<tr>
<td>700</td>
<td>800</td>
<td>135</td>
<td>0.15</td>
</tr>
</tbody>
</table>

Table 1. $VQM_G$ for the watermarked video.
Table 2. Performance of the proposed video watermarking scheme for different video attacks.

<table>
<thead>
<tr>
<th>ATTACK</th>
<th>Proposed approach</th>
<th>Approach in [14]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type I Collusion</td>
<td>0.89</td>
<td>0.79</td>
</tr>
<tr>
<td>Type II Collusion</td>
<td>0.90</td>
<td>0.82</td>
</tr>
<tr>
<td>MPEG-2 (2500 Kb/s) → MPEG-4 (600 Kb/s)</td>
<td>0.92</td>
<td>0.12</td>
</tr>
<tr>
<td>MPEG-2 (1000 Kb/s) → MPEG-4 (600 Kb/s)</td>
<td>0.90</td>
<td>0.10</td>
</tr>
<tr>
<td>Drop 1 frame every 5</td>
<td>0.54</td>
<td>0.16</td>
</tr>
</tbody>
</table>

In this paper we have proposed an informed watermarking method making use of the quantization index modulation approach operating in the 3D-DCT domain. A perceptual criterion based on a qualitative analysis of the HVS has also been employed to compromise between robustness and imperceptibility. In summary, the experimental results have outlined that proposed method exhibits high robustness against most video attacks still maintaining an excellent perceptual invisibility.

6. ACKNOWLEDGMENTS

The authors would like to thank G. Angiolella and D. Interligi for their help in computer simulations.

7. REFERENCES


