Robust 3D DFT Video Watermarking

Deguillaume, Frédéric; Csurka, Gabriela Otilia; O'Ruanaidh, Joséph John; Pun, Thierry

How to cite

DEGUILLAUME, Frédéric et al. Robust 3D DFT Video Watermarking. In: IS\&T/SPIE's 11th Annual Symposium, Electronic Imaging '99: Security and Watermarking of Multimedia Contents III. Ping Wah Wong and Edward J. Delp (Ed.). San Jose (USA). [s.l.] : [s.n.], 1999. p. 113–124. (SPIE Proceedings) doi: 10.1117/12.344662

This publication URL: https://archive-ouverte.unige.ch/unige:47744
Publication DOI: 10.1117/12.344662

Robust 3D DFT Video Watermarking

Frédéric Deguillaume Gabriela Csurka Joseph O'Ruanaidh† Thierry Pun
CUI, University of Geneva, 24 rue Général Dufour, CH 1211 Geneva, Switzerland‡

ABSTRACT

This paper proposes a new approach for digital watermarking and secure copyright protection of videos, the principal aim being to discourage illicit copying and distribution of copyrighted material. The method presented here is based on the discrete Fourier transform (DFT) of three dimensional chunks of video scene, in contrast with previous works on video watermarking where each video frame was marked separately, or where only intra-frame or motion compensation parameters were marked in MPEG compressed videos.

Two kinds of information are hidden in the video: a watermark and a template. Both are encoded using an owner key to ensure the system security and are embedded in the 3D DFT magnitude of video chunks. The watermark is a copyright information encoded in the form of a spread spectrum signal. The template is a key based grid and is used to detect and invert the effect of frame-rate changes, aspect-ratio modification and rescaling of frames. The template search and matching is performed in the log-log-log map of the 3D DFT magnitude.

The performance of the presented technique is evaluated experimentally and compared with a frame-by-frame 2D DFT watermarking approach.

Keywords: video watermarking, MPEG, 3D discrete Fourier transform, log-polar-log and log-log-log mapping, spread spectrum.

1. INTRODUCTION

Digital media have become common and have increasingly taken over the traditional analog media. There are a great number of technical reasons for favoring digital media. Infrastructure such as computers, printers and high rate digital transmission facilities are becoming inexpensive and widely available. Digital networks also provide an efficient cost-effective means of distributing digital media. The popularity of the World Wide Web has clearly demonstrated the commercial potential of the digital multimedia market and consumers are investing heavily in digital audio, image and video recorders and players.

An increasing number of movies and other video documents are recorded on digital supports for public as well as for professional applications. The development of digital video is more recent than that of other medias because of the large bandwidth required. Electronic components however continue growing more powerful, while their cost decrease rapidly. Now DVD is coming on the market, which allows anyone to view movies at home with high quality; its aim is to replace old analog video tape decks. A number of High Definition Television (HDTV) channels are already available thanks to satellites around the world. Soon, interactive Video on Demand (VOD) applications will be proposed to the public where people will be able to individually select movies. In order to facilitate handling and transfer speed, these applications need efficient video compression. The most commonly used compression standard is MPEG-2, which allows high bit-rate (up to 50 Mb/s) and high visual quality; pilot applications based on MPEG-4 are already appearing.

These developments unfortunately afford virtually unprecedented opportunities to pirate copyrighted material. Digital storage and transmission make it trivial to quickly and inexpensively construct *exact* copies. For digital video, compliant readers and recorders are being designed to handle reproduction based on specific headers in the encoded movie containing control bits to authorize copying of the data. Sony uses such a mechanism for its audio Mini-disk, called SCMS* (Serial Copy Management System), to prevent the copy of a record which has already been copied once from the original. This kind of protection however can easily be circumvented by removing the headers; also such headers do not survive format transcoding or partial copying of the data (spatial or temporal cropping). Therefore, to avoid such weaknesses of copyright systems, the trend is to embed copyright information as a digital watermark

† Current address: Siemens Corporate Research, 755 College Road East, Princeton, NJ 08540, US, oruanaidh@scr.siemens.com
‡ Email: {Frederic.Deguillaume,Gabriela.Csurka,Thierry.Pun}@cui.unige.ch; http://cuiwww.unige.ch/~vision
*Digital Audio Interface, International Standard IEC 958.

within the data itself. In this case the control bits and owner signatures cannot be removed easily or at all without destroying the data (except by the owner of the key used to embed the information). Due to this property, digital watermarking has stimulated significant interest and has recently become a very active area of research.

1.1. Paper organization

Section **2** overviews requirements of, and techniques used for image and video watermarking. Section **3** introduces the new concept of 3D DFT watermarking and lists some relevant properties of the 3D Fourier transform. In section **4** the watermarking encoding/decoding and embedding/extraction processes are detailed. Section **5** presents the 3D template matching technique and finally in section **6** experimental results are presented.

2. IMAGE AND VIDEO WATERMARKING

In order to be an efficient part of a reliable and secure image or video copyright protection system, a watermarking algorithm must fulfill the following requirements:

- the watermark has to be secure: an unauthorized detection or removal of the watermark must be impossible;
- the watermark approach should be *oblivious*, i.e. it should be possible to detect and decode the watermark without the use of the original image or video;
- the watermark has to be perceptually invisible and should not degrade the image or video quality;
- the watermark detection must be reliable, with no false detection and if possible no false rejection;
- the watermark has to be resistant to manipulations such as photometric transformations (e.g. filtering or luminance correction), geometric transformations (e.g. translation, rotation, scaling, cropping or aspect ratio change), analog $\leftrightarrow$ digital conversions, scanning, lossy compression.
- the watermark has to resist to cryptographic attacks and other intentional watermark destruction attacks such as the jitter attack,[14] Stirmark,[14] mosaic attack[14] or statistical averaging attack[2];

The following requirements are specific to video watermarking:

- the video watermark has to resist to changes of frame-rate;
- the watermark has to be detectable anywhere in the movie and within a short time (no more than a few seconds);
- the watermark should resist to different video compression schemes as well as to re-compression in a different format;
- the embedding and extraction processes should be performed in real-time, to allow SCMS-like features on compliant recording devices.

Techniques for hiding watermarks in still images have grown steadily more sophisticated and increasingly robust to lossy image compression and standard image processing operations, as well as to cryptographic attack. Many of the current techniques for embedding marks in digital images, inspired by methods of image coding and compression, work in the frequency transformed domain (DCT,[10,1] DFT,[12] Wavelets,[10] Fractals,[17] etc.). One of the key element to make a watermark robust is to embed it in the *perceptually significant* components of the image. The term "perceptually significant" is somewhat subjective but it suggests that a good watermark is one which takes account of the behavior of human visual system.[10,1,3] Another key element is the use of spread spectrum techniques to encode the information before embedding in the images.[18,1] The advantages of a spread spectrum system is that it transforms the narrow band data sequence into a noise-like wide-band signal, using pseudo-random sequences that are difficult to detect and extract.

In the case of video watermarking the challenge is to mark a group of images which are strongly intercorrelated and often manipulated in a compressed form, e.g. MPEG. A first group of video watermarking methods therefore directly operate on MPEG data to avoid full decompression. Hartung *et al.*[7] proposed to mark only the DCT

coefficients of the intra-frames (I-frames). They use a spread spectrum signal containing the copyright information which is added to the non-zero DCT coefficients under the condition of not increasing the bit rate. Other researchers watermark MPEG-2 motion compensation vectors[6] or MPEG-4 facial animation parameters.[5] The advantage of these methods is their rapidity as they need not decompression of the MPEG data. They are, however, not resistant to various transformations of image frames such as rescaling, change of frame-rate, compression and re-compression in a different format or a different GOP (group of pictures) organization.

In order for the watermark to be able to restist such transformations as well as to be less dependent on the way the video compression was done, the basic approach adopted here is to mark the uncompressed video sequence in spite of the increased computational cost. Working with uncompressed video, a first possibility is to individually mark all the frames of the video using a still image watermarking technique. Doing so would allow to inherit the robustness of the 2D approaches; the drawback however would be the vulnerability to averaging attacks, where consecutive frames are averaged to remove the embedded mark.[2]

A second possibility is to take into account the temporal dimension of the video. Hartung *et al.* consider the whole video as a one dimensional signal acquired by line scanning and they embed a watermark in form of spread spectrum into the direct domain of the video stream.[7] Swanson *et al.*[9] propose to mark the static and dynamic temporal components generated from a temporal wavelet transform of the video.

In the method proposed here, in contrast to the former ones, the video is considered as a three-dimensional signal with two dimensions in space and one dimension in time. The basic idea is to extend the two dimensional robust DFT image watermarking scheme described in[11,12] to a three-dimensional DFT video watermarking scheme. With this novel approach, the watermark is embedded into the magnitude of the 3D Discrete Fourier Transform (DFT) of the video data. The ownership and copyright information are encoded in a spread spectrum signal generated by m-sequences or Gold Codes, which is added into the magnitude values of the three-dimensional Fourier transform domain. In addition to the watermark, a *three-dimensional* template-grid is embedded into the 3D DFT magnitude in order to determine and invert geometric transformations suffered by the video. The proposed method is able to determine the parameters of transformations such as frame cropping, frame padding, frame scaling and/or aspect-ratio changes as well as frame-rate changes.

3. SPATIO-TEMPORAL VIDEO WATERMARKING

This section gives a general overview of the 3D DFT spatio-temporal watermarking approach. The proposed method works with uncompressed video data and is independent of the video format encoding. Therefore, in the case of MPEG compressed videos the sequence is decompressed first before embedding or extracting the mark, and re-compressed if necessary, after the watermarking process. Despite the added computational cost this method was motivated by the fact that it verifies most of the requirements listed in section **2**. Indeed, experiments show (see section **6**) that the embedded mark resists to frame cropping and padding, video frame re-sampling, aspect-ratio modification and also MPEG compression in standard quality despite the use of motion compensation. The presented method is *oblivious*, i.e. it does not need any information from the original video during the watermark extraction. Moreover, the method is secure, the system security being insured by the proprietary knowledge of certified, authenticated and securely distributed keys[8] needed to generate the spread spectrum signal.

3.1. 3D-blocks based watermarking

To perform the spatio-temporal based watermarking, the video is viewed as a three-dimensional signal with two dimensions in space and one dimension in time. As performing a three-dimensional DFT of the whole video would be very costly, the video is first divided into consecutive chunks of fixed length (a fixed number of frames, typically 16 or 32 corresponding to approximately 0.5 to 1 seconds of the video scene). In this approach, the video blocks are considered consecutive and non-overlapped for the computation of the Fourier transform. It should however be possible to work with overlapped blocks using a 3D extension of the LOT watermarking technique presented in.[13] The watermark embedding or extraction process is then repeatedly performed within each 3D block independently and the same watermark information is embedded into each block. In order to facilitate the description of the algorithm, a brief reminder of the 3D DFT and its properties is presented.

3.2. 3D Discrete Fourier Transform

Let the 3D video block be considered as a real valued continuous function $f(x, y, z)$ defined on an integer-valued Cartesian grid of size $N_x \times N_y \times N_z$, with $0 \leq x < N_x$, $0 \leq y < N_y$, and $0 \leq z < N_z$. Its Discrete Fourier Transform (DFT) in 3D is defined as follows:

$$F(k_x, k_y, k_z) = \sum_{x=0}^{N_x-1} \sum_{y=0}^{N_y-1} \sum_{z=0}^{N_z-1} f(x, y, z) e^{-j2\pi x k_x/N_x - j2\pi y k_y/N_y - j2\pi z k_z/N_z} \tag{1}$$

The inverse transform is given by:

$$f(x, y, z) = \frac{1}{N_x N_y N_z} \sum_{k_x=0}^{N_x-1} \sum_{k_y=0}^{N_y-1} \sum_{k_z=0}^{N_z-1} F(k_x, k_y, k_z) e^{j2\pi k_x x/N_x + j2\pi k_y y/N_y + j2\pi k_z z/N_z} \tag{2}$$

The DFT of a real block is generally complex valued. This leads to a magnitude and a phase representation:

$$A(k_x, k_y, k_z) = |F(k_x, k_y, k_z)| = \sqrt{Re(F(k_x, k_y, k_z))^2 + Im(F(k_x, k_y, k_z))^2}$$
$$\Phi(k_x, k_y, k_z) = \angle F(k_x, k_y, k_z) = arctan\Big(\frac{Im(F(k_x, k_y, k_z))}{Re(F(k_x, k_y, k_z))}\Big)$$

Note that for the resulting $F(k_x, k_y, k_z)$, the (k_x, k_y) plane corresponds to the spatial frequencies of frames, while the k_z axis represents the temporal frequencies. Moreover, the DFT is a separable function, i.e. it corresponds to three one-dimensional DFT applied consecutively along the three axes over the entire data set.

3.3. Fourier Transform properties

Amongst the numerous properties of the DFT,[16] only those which are relevant to the present algorithm, are presented. The properties listed here concern the effects on the DFT of linear transformations. They are presented in 3D, however they are valid for any dimension.

1. Scaling:

   Consider a scaling with independent scaling factors in each direction of the 3D block and apply the DFT on the rescaled block. The DFT will be equally rescaled, with inverse scaling factors along the axes in the frequency domain:

   $$F(\frac{k_x}{\lambda_x}, \frac{k_y}{\lambda_y}, \frac{k_z}{\lambda_z}) \leftrightarrow f(\lambda_x x, \lambda_y y, \lambda_z z) \tag{3}$$

   Note that for 3D video blocks the λ_x and λ_y scaling factors correspond to changes of proportions inside each frame, while the λ_z factor represents the temporal scaling, i.e. the change of the video frame-rate.

2. Rotation:

   Consider a 3D rotation of the block in the (x, y) plane around the z axis. Rotating all frames with an angle θ in the direct domain causes a rotation with the same angle θ in the (k_x, k_y) plane and hence a global rotation around the k_z axis of the DFT in the frequency domain:

   $$\begin{aligned} &F(k_x \cos\theta - k_y \sin\theta, k_x \sin\theta + k_y \cos\theta, k_z) \\ &\quad \leftrightarrow f(x \cos\theta - y \sin\theta, x \sin\theta + y \cos\theta, z) \end{aligned} \tag{4}$$

   In fact this property is true for an arbitrary rotation in space. For video sequences, however, it is very unlikely that the video scene will be rotated in another plane in spatio-temporal space.

3. Translation:

   Shifting the video block in the direct domain causes a linear shift in the frequency domain:

$$F(k_x, k_y, k_z) \exp\left[-j(ak_x + bk_y + ck_z)\right] \leftrightarrow f(x + a, y + b, z + c) \tag{5}$$

   Note that both $F(k_x, k_y, k_z)$ and its dual $f(x, y, z)$ are considered by the DFT as periodic functions and therefore it is implicitly assumed that translations corresponds to a "wrap around", i.e . the part of the video or DFT getting out of the considered block in a given direction reappears on the other side. This can be referred to as a *circular translation* or cyclic shift, and due to this property the Fourier transform is circular translational invariant.

4. THE WATERMARK

Two types of information are embedded in the 3D DFT magnitude of video blocks: a watermark and a template (described in detail in the next section). The watermark is a message, that can contain information such as the owner of the image, a serial number, flags indicating copying information, type of content, etc., or alternatively a hash number to a table that contains these elements of information. It is encoded using pseudo-random sequences, such as m-sequences or Gold Codes to obtain a spread spectrum signal which will be embedded in the 3D DFT magnitude.

4.1. Watermark encoding

The encoding process is briefly presented here, more details concerning the properties of the spread spectrum systems, m-sequences or Gold Codes can be found in the e.g.[15,4] The basic idea is to represent the message in binary form $\boldsymbol{b} = (b_1, b_2, ...b_M)^\top$, with $b_i \in \{1, -1\}$. For each bit b_i a pseudo-random sequence $\boldsymbol{v}_i$ is generated. Here the family $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_M\}$ are shifted m-sequences or a family of Gold Codes due to their favorable statistical and cryptographic security properties.

The encoded message can therefore be obtained as follows:

$$\boldsymbol{w} = \sum_{i=1}^{M} b_i \boldsymbol{v}_i = \mathbf{G}\boldsymbol{b} \tag{6}$$

where $\boldsymbol{b}$ is a $M \times 1$ vector of bits and $\mathbf{G}$ is an $N \times M$ matrix such that the i^{th} column of $\mathbf{G}$ is a pseudo-random vector $\boldsymbol{v}_i$, all coefficients being ± 1. The resulting $N \times 1$ vector $\boldsymbol{w}$ is the watermark in the form of a spread spectrum signal which has to be embedded in the 3D DFT magnitude.

4.2. Watermark embedding

In order to add the values of the spread spectrum sequence $\boldsymbol{w}$ to the components of the DFT's magnitude, the 2D approach presented in[11,12] is generalized. In this 2D approach the watermark is added to the 2D DFT magnitude inside of a medium frequency band. The reason is the desired compromise between visibility and robustness to lossy compression. Indeed, on the one hand, the largest part of the image energy is concentrated in low frequencies, therefore modifying them may result in fairly visible artifacts. On the other hand, high frequencies are often easily removed through lossy compression and therefore should be avoided.

As the 3D DFT has the separability property, i.e. it corresponds to a 2D DFT in spatial domain followed by a 1D DFT along the temporal axis, the property concerning the spatial frequencies range remains valid. Therefore, similarly to the 2D DFT, a mid-range frequency band is chosen in the spatial domain. Let r_{min} be the lower and r_{max} be the upper frequency bounds in the DFT domain ($0 < r_{min} < r_{max}$). Further, for the video blocks, the third dimension corresponding to the temporal frequencies also has to be taken into account. These frequencies have the following properties. Null or low temporal frequencies are linked to static components in the input scene, while higher frequencies are related to moving objects and varying areas. Therefore again, due this time to a compromise between the static and moving components, a mid range is considered. Let d_{min} and d_{max} be the lower respectively temporal frequency bound.

Assume that the center or origin of the DFT (null frequency point) is in the middle of the block (if not a DFT shift can be performed). Then, the Fourier transform is symmetric, the center of the symmetry being the center of

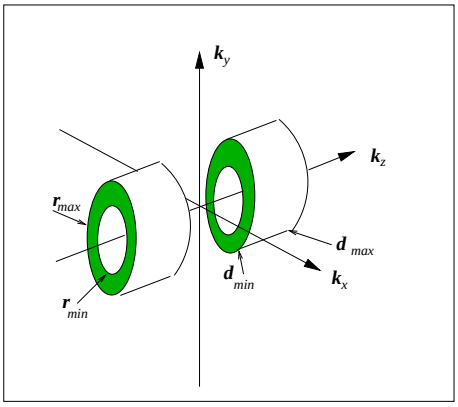

Figure 1. *Embedding area of the watermark in the 3D DFT's magnitude blocks.*

the DFT. Consequently, the volume considered for marking comprise two cylindric annuli in the DFT block as shown in Figure 1.

The spread spectrum signal is added to the magnitude values of this annulus. However, the DFT magnitude has to be kept symmetric relative to the origin, otherwise the inverse DFT will have non-real values. For this reason, only one of the two cylindric annulus is considered and each time a magnitude value at the position (k_x, k_y, k_z) is modified, the same value is added to the magnitude at the position $(-k_x, -k_y, -k_z)$.

Moreover, the generated watermark $\boldsymbol{w}$ contains positive and negative values (it is easy to see from (6) that $w_i \in \{-M, M\}$). Adding w_i to positive definite magnitude values can lead to negative values and the DFT will have non-real values. To avoid this, a new representation of positive and negative values is introduced. The basic idea is to replace each value w_i of the sequence $\boldsymbol{w}$ by the pair:

$$w_i \rightarrow \boldsymbol{p}_i = \begin{cases} (w_i, 0) & if \quad w_i \geq 0 \\ (0, -w_i) & if \quad w_i < 0 \end{cases} \tag{7}$$

Using this representation and choosing a pair of magnitude values for each $\boldsymbol{p}_i$, the first value is modified (by adding w_i) if $w_i \geq 0$ and the second one is modified (by adding $|w_i|$) if $w_i < 0$. The DFT locations for the pair of values can be chosen arbitrarily, the only constrain being to use the same pair-wise arrangements during the embedding and the extraction.

Consequently, the maximum number of spread spectrum values that can be embedded in the above described manner corresponds to the number of pixels contained in a half of the cylindric annulus (a quarter of the whole marked area, as it is divided by two because of the symmetry and divided again by two because of the pair-wise embedding of w_i). However, the large amount of space available in 3D blocks of videos widely compensates for this drawback, especially compared to 2D watermarking of images.

4.3. Watermark extraction

In order to extract the watermark, the compressed video is uncompressed first and the video sequence is divided into consecutive fixed length blocks. Due to the shift invariant and wrap-around property of the DFT, it is not necessary to have an exact synchronization of the 3D blocks according to the embedded steps. Note that this is true only if the embedded marked is the same in each 3D video chunk.

Then each 3D block is DFT transformed and the template search and matching, described in section **5** in order to retrieve the transformations applied to the video, is performed. Using the parameters of the transformation the new positions of the pair-wise magnitudes can be computed. The difference between the pair-wise coefficients allows to obtain a spread spectrum signal:

$$\boldsymbol{w}' = \boldsymbol{w} + \boldsymbol{e}$$

where $\boldsymbol{w}$ is the embedded watermark and the error $\boldsymbol{e}$ contains the image and/or additive noise. Note that the two magnitude values of the pair-wise points chosen during the embedding process are not necessary equal, which means that even if the extraction is done on the marked video on which no manipulation was done, $\boldsymbol{e}$ is not a zero vector.

4.4. The watermark decoding

It can be shown that for m-sequences and Gold Codes the dot product between $\boldsymbol{v}_i$ and $\boldsymbol{v}_j$ is:

$$< \boldsymbol{v}_i, \boldsymbol{v}_j >= \begin{cases} N & \text{if} \quad i = j \\ -1 & \text{if} \quad i \neq j \end{cases} \tag{8}$$

Therefore, in order to decode the message from $\boldsymbol{w}'$, for each i the dot product ("cross-correlation") between $\boldsymbol{v}_i$ and

$$\boldsymbol{w}' = \sum_{i=1}^{M} b_i \boldsymbol{v}_i + \boldsymbol{e}$$

is performed. Hence,

$$B'_j =< \boldsymbol{w}', \boldsymbol{v}_j >= \sum_{i=1}^{M} b_i < \boldsymbol{v}_i, \boldsymbol{v}_j > + < \boldsymbol{e}, \boldsymbol{v}_j >$$

and replacing (8) gives:

$$B'_j = b_i\, N - (M-1) + < \boldsymbol{e}, \boldsymbol{v}_j >$$

Generally $M \ll N$. Moreover the distribution of $\boldsymbol{e}$ can be approximated by a normal distribution with zero mean, so $< \boldsymbol{e}, \boldsymbol{v}_j >$ is negligible comparing to N. Therefore, each embedded information bit b_j can be retrieved as follows:

$$b'_j = sign(B'_j) = sign(b_j) = b_j$$

5. THE TEMPLATE

The watermarking method described in the previous section presents an inherent invariance to spatial shifts or temporal shift of the sequence due to the basic properties of the Fourier transform. It also resists to simple filtering, noise adding, MPEG compression, format re-encoding, etc., because the spread-spectrum sequences are very robust to noise or partial cancelation. However, the watermark without the template does not resist to other transformation such as frame cropping, frame scaling or changes of aspect ratio, frame rotations or changes of frame-rate. This is because these transformations modify the position of the embedded signal inside the DFT magnitude and spread-spectrum sequences need perfect synchronization to work properly. Retrieving these new positions is equivalent to recovering the geometrical transformation applied to the video. To be able to estimate the parameters of these transformations a template is inserted into the 3D DFT magnitude in addition to the watermark.

The template is a sparse set of points embedded symmetrically to preserve the symmetry property of the Fourier transform. Another requirement that a template point has to verify is to be a local maximum (peak) in the DFT magnitude in order to facilitate the template search process (see section **5.3**).

The template is a key based grid and therefore it is possible to create a reference template during the extraction. This reference template is then used to search for the transformed template in the 3D DFT magnitude of the modified video. An exhaustive search in the DFT domain would be very costly as the transformation is unknown. In order to circumvent this problem, a log-type mapping is used which allows to transform the scaling or rotation into simple shift operations. Two basic mappings are considered here corresponding to two main video manipulations. The first is the log-log-log map which is used to retrieve the rescaling of the video including the frame scaling and aspect ratio change, as well as the frame rate changes. The second is the log-polar-log map allowing to recover a less usual transformation that consists of frame rate changes combined with a rotation and scaling applied to each frame.

5.1. The 3D log-log-log map

The "log-log-log" map allows to find independent scalings along the three axes corresponding to a aspect-ratio change and a frame-rate change. It is a bijective function (if the origin $(0,0,0)$ was extracted) and it converts the DFT (k_x, k_y, k_z) space to a (μ_x, μ_y, μ_z) logarithmic space as follows:

$$\begin{array}{l} \mu_x = sign(k_x) \cdot ln(|k_x|) \\ \mu_z = sign(k_y) \cdot ln(|k_y|) \\ \mu_z = sign(k_z) \cdot ln(|k_z|) \end{array} \tag{9}$$

Note that k_i must be not equal to zero, which must be taken into account during the generation of the template positions.

From the signed shifts $\triangle\mu_x$, $\triangle\mu_y$ and $\triangle\mu_z$ in the log-log-log space obtained by the template matching (section **5.3**) the rescaling factors s_x, s_y, s_z of the 3D DFT[†] are as follows:

$$s_x = e^{\triangle\mu_x} \quad , \quad s_y = e^{\triangle\mu_y} \quad , \quad s_z = e^{\triangle\mu_z} \tag{10}$$

The new positions in the DFT space of the watermark is given by:

$$\begin{pmatrix} \hat{k}_x \\ \hat{k}_y \\ \hat{k}_z \end{pmatrix} = \begin{bmatrix} s_x & 0 & 0 \\ 0 & s_y & 0 \\ 0 & 0 & s_z \end{bmatrix} \begin{pmatrix} k_x \\ k_y \\ k_z \end{pmatrix} \tag{11}$$

where (k_x, k_y, k_z) are the positions where the mark was embedded.

5.2. The 3D log-polar-log map

The "log-polar-log" map allows to find a rotation and a uniform scaling of frames (keeping the same aspect-ratio), as well as a change of frame-rate. It is a bijective function (again if the origin $(0,0,0)$ was extracted) and converts the (k_x, k_y, k_z) space to a (ρ, θ, μ_z) circular and logarithmic space as follows:

$$\begin{array}{l} \rho = ln\left(\sqrt{{k_x}^2 + {k_y}^2}\right) \\ \theta = arctan\left(\frac{k_y}{k_x}\right) \\ \mu_z = sign(k_z) \cdot ln(|k_z|) \end{array} \tag{12}$$

where $\theta \in [0, 2\pi[$ and $k_z \neq 0$. Given the shifts $\triangle\rho$, $\triangle\theta$ and $\triangle\mu_z$ obtained by template matching, the spatial scaling s, the rotation angle δ and the temporal resampling s_z are obtained from:

$$s = e^{\triangle\rho} \quad , \quad \delta = \triangle\theta \quad , \quad s_z = e^{\triangle\mu_z} \tag{13}$$

Finally, the compensated positions are given by:

$$\begin{pmatrix} \hat{k}_x \\ \hat{k}_y \\ \hat{k}_z \end{pmatrix} = \begin{bmatrix} s \cdot \cos\delta & -s \cdot \sin\delta & 0 \\ s \cdot \sin\delta & s \cdot \cos\delta & 0 \\ 0 & 0 & s_z \end{bmatrix} \begin{pmatrix} k_x \\ k_y \\ k_z \end{pmatrix} \tag{14}$$

Unfortunately, neither using the log-log-log map nor using the log-polar-log map allows to retrieve both the aspect-ratio changes and the rotation of frames. It is however unlikely that video frames be rotated, and therefore the experimental results (section **6**) are focused only on the log-log-log mapping.

[†] s_x, s_y, s_z correspond to the inverse scaling of the video in accordance to (3).

5.3. Efficient template search

It was shown that mapping the DFT to a log-type domain transforms the scaling or rotation to a simple 3D shift in the log-log-log respectively log-polar-log space. This means that in order to retrieve the transformation a cross-correlation step needs to be applied between the reference template and the mapped DFT. This is a nice solution in theory, which however can be computationally very costly. To avoid this a different approach is proposed here based on the following properties. The template is made of a sparse set of points, and the transformed template points can only be found amongst local peaks in the DFT magnitude as was ensured during the embedding. The following algorithm takes into account these properties and therefore allows for an efficient and low cost template matching:

- from the reference template only the list of the point coordinates are kept. They are noted by $\boldsymbol{p}_i = (k_x, k_y, k_z)_i$, where $i = 1..n_t$, n_t being the number of embedded template points;
- in the DFT magnitude domain all the local maxima (or peaks) are extracted. The positions of the extracted peaks are denoted $\boldsymbol{p}'_j = (k'_x, k'_y, k'_z)_j$, $j = 1..m_t$. Usually, $m_t \gg n_t$ as other peaks will appear in addition to the embedded template points;
- coordinates of both lists are converted to the desired log-type mapping using equations (9) or (12). The modified coordinates are denoted by $\widehat{\boldsymbol{p}}_i = (\hat{k}_x, \hat{k}_y, \hat{k}_z)_i$ for the reference template and by $\widehat{\boldsymbol{p}'}_j = (\hat{k}'_x, \hat{k}'_y, \hat{k}'_z)_j$ for the peaks set;
- in order to retrieve the global shift, the two sets of points are matched as follows. For each pair of points $(\widehat{\boldsymbol{p}}_i, \widehat{\boldsymbol{p}}'_j)$ the translational vector $\boldsymbol{v}_{ji} = \widehat{\boldsymbol{p}}'_j - \widehat{\boldsymbol{p}}_i$ is computed and added to each reference point $\widehat{\boldsymbol{p}}_k$. This means that all reference point are virtually shifted to a new position according to the vector $\boldsymbol{v}_{ji}$. Further, for each reference point $\widehat{\boldsymbol{p}}_k$ the distance between its translated position and the peak which is the closest to this new position is computed as follows:

$$d_{ij}^k = \min_l \{(\widehat{\boldsymbol{p}}_k + \widehat{\boldsymbol{v}}_{ji}) - \widehat{\boldsymbol{p}}'_l\}$$

  Finally, the 3D shift $\boldsymbol{v}_{ji}$ that minimizes the sum of distances:

$$\min_{i,j} \{\sum_{k=1}^{n_t} d_{ij}^k\}$$

  is retained. It corresponds to the global 3D shift of the template and allows to compute the scaling or rotation parameters by using the equations (10) or (13).

The exhaustive template search has a complexity of ${n_t}^2 \cdot {m_t}^2$ which is low relatively to a cross-correlation approach. This complexity can even be further decreased. Indeed, in theory, the template search works even if instead of all pairs $(\widehat{\boldsymbol{p}}_i, \widehat{\boldsymbol{p}}'_j)$ only the pairs with a fixed template point $\widehat{\boldsymbol{p}}_i$ are used. The only condition is that this template point lies between the peaks, which cannot always be ensured in practice. However, a subset of shifts $\boldsymbol{v}_{ji}$ with $i \in I_0 \subset \{1..n_t\}$ and $j = 1..m_t$ will ensure that the correct shift was considered while decreasing the complexity.

Another gain obtained by using this approach is in the increased precision of the extracted transformation parameters, since all computations of log-type mappings and template search make use of floating point coordinates.

6. EXPERIMENTAL RESULTS

In this section some results are presented which show the robustness of the 3D DFT watermarking to MPEG compression, changes of proportions and temporal resampling. For the experiments two MPEG-2 sequences were used, which have been encoded from PAL television images with a frame-rate of 25 frames/sec. The frames were of size 352×288 pixels, which is close to the 4/3 TV standard format ($352/288 = 3.667/3$). The first sequence, called "no comment", was 128 frames long (about 5 s), and the second one, called "news", was 256 frames long (10 s). Both video sequence were decompressed before performing any watermarking operation.

The 3D DFT algorithm as described in this paper was first tested. The efficient template matching was applied in the log-log-log map of the 3D DFT. For the "no comment" sequence, blocks of 16 frames and for the "news" sequence,

Figure 2. *An original and a 3D DFT watermarked frame. Left: "no comment" sequence. Right: "news"sequence .*

blocks of 32 frames were used (see Figure 2). Note that the embedded mark verify the invisibility requirement as shown in Figure 2.

In order to compare the results of the approach with a frame-by-frame one, the 2D DFT watermarking method described in[12] was performed on each frame. The 2D template matching was achieved in the log-log rather than log-polar map of the 2D DFT magnitude since aspect ratio changes are more likely to occur than rotations. Note that this single frame watermarking can also be seen as a 3D DFT scheme where the length of the blocks is considered equal to 1 frame.

For both sequences and both algorithms the embedded message, "Hello Earth", was of 88 bits long. The main tests concerned the resistance to MPEG compression, aspect-ratio changes with or without additional recropping/padding of frames to their original size (see Figure 3). To test the resistance to MPEG-2 the recompression was done at the "main" profile, "main" level and in standard quality mode.

Figure 3. *Movies rescaled from 4/3 to 2.11/1. Left: rescaled "no comment". Center: rescaled "news", cropped to its original size. Right: rescaled "news", padded to its original size.*

For the 3D DFT approach, the resistance to changes of frame-rate and desynchronization of 3D-blocks were also tested. Desynchronization means that during the extraction process, the 3D-blocks were not synchronized in time with the 3D-blocks used during the embedding process. Note that 4/3, 16/9 and 2.11/1 are standard aspect-ratios, while 23.976, 25 and 30 frames/sec (fps) are standard frame-rates.

The results of the 3D DFT watermarking method are summarized in Table 1, and those of the frame-by-frame approach in Table 2. Each time a test is performed, the bit error rate (BER), i.e. the proportion of wrong bits in the decoded message, is computed. Three BER classes were considered, the first class corresponding to no error, the second class containing 5% or less of wrong bits (corresponding to 1-4 bit over 88 and considered recoverable by error-control coding) and finally the third class containing results where more than 5% of bits were wrong (considered as a failure). Note that no error control coding was implemented in the current version of the algorithm. The tables (1 and 2) present the percentage of 3D-blocks, respectively frames, which fell into each classes.

Comparing the results of the two approaches, one can see that they are very similar. Both methods have a good robustness to MPEG compression and to common frame transformations such as scaling or aspect-ratio change. In addition, the 3D approach shows a good resistance to desynchronization due to the Fourier transform shift invariance property. Finally, experimental results confirm that due to the efficient template matching algorithm the 3D DFT

operation	ber 0	ber 5%	failed
MPEG-2 compression only	100%	0%	0%
MPEG-2 + desynchronization	86%	7%	7%
MPEG-2 + aspect-ratio changes			
4/3 vs. 16/9	94%	6%	0%
4/3 vs. 2.11/1	92%	2%	6%
MPEG-2 + frame-rate changes			
25 vs. 30 fps.	92%	8%	0%
23.976 vs. 30 fps.	78%	18%	4%
MPEG-2 + aspect-ratio + frame-rate changes			
4/3 vs. 16/9, 25 vs. 30 fps.	87%	11%	2%
4/3 vs. 2.11/1, 23.976 vs. 30 fps.	74%	13%	13%

Table 1. *Results for the 3D video watermarking approach. The test have been made using blocks of 16 and 32 frames long, embedding the same watermark and the same template into each block.*

operation	ber 0	ber 5%	failed
MPEG-2 compression only	100%	0%	0%
MPEG-2 + changes of aspect ratio			
4/3 vs. 16/9	89%	8%	3%
4/3 vs. 2.11/1	85%	13%	2%

Table 2. *Results for the 2D video watermarking, where each frame was marked separately.*

method can easily handle the frame rate changes and detect the watermark (see Table 1). The 2D scheme is obviously not affected by synchronization or frame rates changes, as the same mark was embedded independently in each frame.

The advantage of the 3D approach compared to the frame-by-frame approach is first that it is more robust to averaging attacks since the mark is spread into a volume taking into account the temporal dimension of the video. In addition, the 3D approach offers a larger bandwidth to hide data.

7. CONCLUSION

A new oblivious approach has been presented for video watermarking which, in contrast to existing methods, considers the video as a three-dimensional signal with two dimensions in space and one dimension in time, and embeds the watermark in the 3D DFT of three dimensional chunks of video scene. In addition to the watermark, a three-dimensional template-grid is also embedded into the 3D DFT magnitude, in order to determine possible geometric transformations suffered by the video. To perform the template matching an efficient and low cost algorithm has also been proposed.

The experiments show that the proposed method is able to discover the parameters of the transformation for frames cropping, padding, scaling and/or aspect-ratio changes as well as for frame-rate changes. To retrieve the mark it is not necessary during extraction to know the size and synchronization of the blocks used during the embedding process. Moreover, the method is also resistant to MPEG compression and verifies the invisibility and security requirements.

The main drawback of the method is that it is time consuming due to the need for 3D DFT and for decompressing each sequence for watermark embedding or extraction. This problem can however be alleviated via the use of cheap chips[‡] for a real-time MPEG compression or real-time 3D FFT. It is also worth note that for other applications such as tracing copyright violations, mark embedding or extraction do not necessarily have to be done on user's device; in such application also robustness is a more important requirement than speed.

[‡]E.g. DVxploreTM proposed by C-Cube Microsystems; http://www.c-cube.com/products/dvxplore.html.

ACKNOWLEDGMENTS

We are grateful Dr Alexander Herrigel and Digital Copyright Technologies Switzerland for their work on the security architecture for the digital watermark and for the ongoing collaboration. This work is financed by the European Esprit Open Microprocessor Initiative (project JEDI-FIRE) and by the Swiss Priority Program on Information and Communication Structures (project Krypict). This work is part of the European Patent application EU978107084. Thanks also to Shelby Pereira for his very useful insights.

REFERENCES

1. I. Cox, J. Killian, T. Leighton, and T. Shamoon. Secure spread spectrum watermarking for images, audio and video. In *Proceedings of the IEEE Int. Conf. on Image Processing ICIP-96*, pages 243–246, Lausanne, Switzerland, September 16-19 1996.
2. I. J. Cox and J.-P. M. G. Linnartz. Some general methods for tampering with watermarks. *IEEE Journal on Selected Areas in Communications*, 16(4):587–593, May 1998.
3. J. F. Delaigle, C. De Vleeschouwer, and B. Macq. Digital Watermarking. In *Conference 2659 - Optical Security and Counterfeit Deterrence Techniques*, San Jose, February 1996. SPIE Electronic Imaging : Science and Technology. pp. 99-110.
4. E.H. Dinan and B. Jabbari. Spreading codes for direct sequence cdma and wideband CDMA cellular network. *IEEE Communications Magazine*, June 1998.
5. P. Eisert F. Hartung and B. Girod. Digital watermarking of MPEG-4 facial animation parameters. *Computers and Graphics*, 22(3), 1998.
6. M. Kutter F. Jordan and T. Ebrahimi. Proposal of a watermarking tehnique for hiding/retrieving data in compressed and decompressed video. Technical report, Technical report M2281, ISO/IEC document, JTC1/SC29/WG11, MPEG-4 meeting, Stockholm, Sweden, July 1997.
7. F. Hartung and B. Girod. Watermarking of uncompressed and compressed video. *Signal Processing*, 66:283–301, 1998.
8. A. Herrigel, J. J. K. Ó Ruanaidh, H. Petersen, S. Pereira, and T. Pun. Secure copyright protection techniques for digital images. In *International Workshop on Information Hiding*, Portland, OR, USA, April 1998.
9. B. Zhu M. D. Swanson and A. H. Tewfik. Multiresolution scene-based video watermarking using perceptual models. *IEEE Journal on Selected Areas in Communications*, 16(4):540–550, May 1998.
10. J. J. K. Ó Ruanaidh, W. J. Dowling, and F. M. Boland. Watermarking digital images for copyright protection. *IEE Proceedings on Vision, Signal and Image Processing*, 143(4):250–256, August 1996.
11. J. J. K. Ó Ruanaidh and T. Pun. Rotation, scale and translation invariant spread spectrum digital image watermarking. *Signal Processing*, 66(3):303–317, May 1998. (Special Issue on Copyright Protection and Control, B. Macq and I. Pitas, eds.).
12. S. Pereira, J. J. K. Ó Ruanaidh, F. Deguillaume, G. Csurka, and T. Pun. Template based recovery of Fourier-based watermarks using Log-polar and Log-log maps. Submitted to the International Conference on Multimedia Computing and Systems, Special Session on Multimedia Data Security and Watermarking, June 7-11, Florence, Italy., 1999.
13. S. Pereira, J. J. K. Ó Ruanaidh, and T. Pun. Secure robust digital image watermarking using the lapped orthogonal transform. In *IS&T/SPIE Electronic Imaging'99, Session: Security and Watermarking of Multimedia Contents*, San Jose, CA, USA, January 1999.
14. F. A. P. Petitcolas and R. J. Anderson. Weaknesses of copyright marking systems. In *Multimedia and Security Workshop at ACM Multimedia'98, Bristol, U.K.*, September 1998.
15. R. L. Pickholtz, D. L. Schilling, and L. B. Milstein. Theory of spread spectrum communications – A tutorial. *IEEE Transactions on Communications*, COM-30(5):855–884, May 1982.
16. J. G. Proakis and D. G. Manolakis. *Introduction to Digital Signal Processing.* Maxwell Macmillan Publishing Company, New York, 1989.
17. J. Puate and F. Jordan. Using fractal compression scheme to embed a digital signature into an image. In *Proceedings of SPIE Photonics East'96 Symposium*, November 1996.
18. A. Z. Tirkel, G. A. Rankin, R. G. van Schyndel, W. J. Ho, N. R. A. Mee, and C. F. Osborne. Electronic watermark. In *Dicta-93*, pages 666–672, Macquarie University, Sydney, December 1993.