EURASIP Journal on Applied Signal Processing 2004:14, 2174–2184

Image Content Authentication Using Pinned Sine Transform

Anthony T. S. Ho
School of Electrical & Electronic Engineering, Nanyang Technological University, Nanyang Avenue, Singapore 639798
Email: etsho@ntu.edu.sg

Xunzhan Zhu
School of Electrical & Electronic Engineering, Nanyang Technological University, Nanyang Avenue, Singapore 639798
Email: xzzhu@pmail.ntu.edu.sg

Yong Liang Guan
School of Electrical & Electronic Engineering, Nanyang Technological University, Nanyang Avenue, Singapore 639798
Email: eylguan@ntu.edu.sg

Received 23 October 2003; Revised 24 December 2003

Digital image content authentication addresses the problem of detecting any illegitimate modification on the content of images. To cope with this problem, a novel semifragile watermarking scheme using the *pinned sine transform* (PST) is presented in this paper. The watermarking system can localize the portions of a watermarked image that have been tampered maliciously with high accuracy as well as approximately recover it. In particular, the watermarking scheme is very sensitive to any texture alteration in the watermarked images. The interblock relationship introduced in the process of PST renders the watermarking scheme resistant to content cutting and pasting attacks. The watermark can still survive slight nonmalicious manipulations, which is desirable in some practical applications such as legal tenders. Simulation results demonstrated that the probability of tamper detection of this authentication scheme is higher than 98%, and it is less sensitive to legitimate image processing operations such as compression than that of the equivalent DCT scheme.

Keywords and phrases: semifragile watermarking, content authentication, pinned sine transform.

1. INTRODUCTION

While digital media offer many distinct advantages over their analog counterparts, the ease with which they can be edited and tampered makes the protection of their integrity and authenticity a serious and important issue. In certain practical applications, such as remote sensing, legal defending, news reporting, and medical archiving, there is a need for verification or authentication of the integrity of the media content. A *fragile watermarking* detects changes of the watermarked image such that it can provide some form of guarantee that the image has not been tampered with and is originated from the right source. In addition, a fragile watermarking scheme should be able to identify which portions of the watermarked data are authentic and which are corrupted; if unauthenticated portions are detected, it should be able to restore it [1].

The earliest fragile watermarking schemes are designed to detect any slight changes to the bits of the watermarked image and the watermark becomes undetectable after the watermarked image is modified in any way [2, 3, 4, 5]. However, since the meaning of multimedia data is generally based on their semantic content rather than the bit streams, in some applications, a *semifragile watermarking* is more desirable. A semifragile watermarking seeks to verify that the content of the multimedia has not been modified by any predefined set of illegitimate distortions, while allowing modification by legitimate distortions [1]. Although a variety of semifragile watermarking schemes have been proposed in the literature to solve this problem, the above issue of "selective content authentication" has not been vigorously addressed.

In [6], Lin and Chang proposed a method that could localize malicious tampering to the image content while accepting JPEG compression to a predetermined quality factor (QF). Their method achieved its goal by using an invariant relationship between two DCT coefficients in a block pair before and after JPEG compressions. Such relationship was encoded and inserted into the least significant bits (LSBs) of rounded DCT coefficients. Although their method proved to

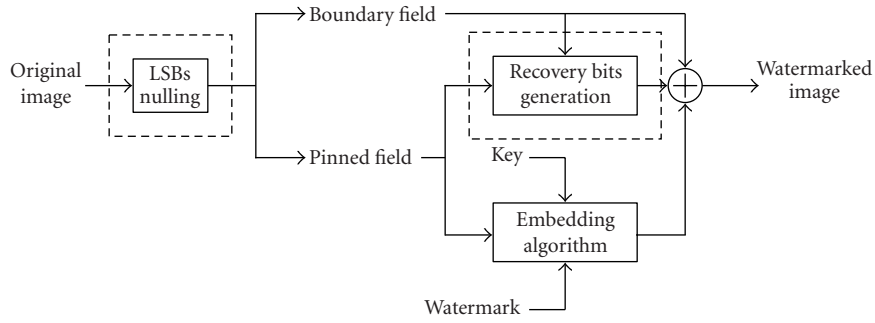


Figure 1: Watermark embedding process; the parts in the dashed windows are optional for the host image restoration.

be robust to JPEG compression by both mathematical deduction and experimental results, they actually proposed a watermarking scheme that was very robust to JPEG compression rather than addressed the issue of selective content authentication. Recently, some fragile watermarking schemes using the wavelet domain have been proposed [7, 8, 9, 10]. The localization ability in both spatial domain and frequency domain makes the wavelets a potential candidate for semifragile watermarking. However, to authenticate content, some significant features, for example, the edges of the host image, are required to be encoded and embedded in the low frequencies of the wavelet decomposition. Thus, there exists a tradeoff between the visual quality of the watermarked image and the ability of the scheme to detect changes. Another drawback of these schemes is the high computation cost during the feature extraction and visual hash coding processes.

Further ways to completely thwart many existing fragile watermarking schemes are the "cutting and pasting" attacks. The well-known vector quantization (VQ) counterfeiting attacks [11] is one of such attacks. Some inter-relationship between the watermarked blocks is introduced to avoid the VQ attacks [4, 5, 6]; however, a close relationship between uncorrelated blocks may come at the cost of reduced error localization properties and introduce confusion for the consequent authentication process.

In this paper, a novel semifragile watermarking scheme using the pinned sine transform (PST) in [12] is proposed. The motivation for developing a semifragile watermarking based on PST is due to the observation that this transform could provide an effective way to solve both the abovementioned selective content authentication problem and the issue of exposing the cutting and pasting counterfeiting attacks. The observation is as follows. The PST conducts a decomposition of the original image into two mutually uncorrelated fields, namely, the boundary field and the pinned field. The texture information of the original image is contained in the pinned field, wherein the sine transform is equivalent to a fast Karhunen-Loeve transform (KLT). By exploiting this important property, we propose to embed a watermark signal into the sine transform domain of the pinned field for content authentication. As illustrated in this paper, the proposed watermarking scheme is especially sensitive to texture alterations of the host image while permitting controlled amount of modifications to nontexture aspects of the host image. Moreover, although our scheme is blockwise, the watermarking of one block is closely related to all the blocks surrounding it, in a way that will become apparent later in this paper, which renders our scheme robust to the cutting and pasting attacks.

Section 2 presents a brief review of the PST. The proposed watermark embedding and image authentication processes are then described in Sections 3 and 4, respectively. In Section 5, we discuss how the proposed scheme ensures a selective content authentication. The proposed scheme's resistance to VQ counterfeiting attacks is demonstrated in Section 6, followed by experimental results and the conclusion in Sections 7 and 8.

2. THE PINNED SINE TRANSFORM

An overview of the PST is discussed in this section. Suppose a data vector

$$\mathbf{X} = \begin{bmatrix} x_0 & \cdots & x_{n+1} \end{bmatrix}^T \tag{1}$$

is separated into a boundary response $\mathbf{X}^b$ defined by x_0 and x_{n+1}, and a residual sequence $\mathbf{X}' - \mathbf{X}^b$, where

$$\mathbf{X}' = \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix}^T. \tag{2}$$

In [13], Jain showed that if $\mathbf{X}$ is a first-order stationary Gauss-Markov sequence, the sequence $\mathbf{X}' - \mathbf{X}^b$ will have the sine transform as its KLT.

Extending the above theory to the more general 2D case, Meiri and Yudilevich [12, 14] proposed the PST for images. An image field is decomposed into two subfields, namely, the boundary field and a residual field. The boundary field depends only on the block boundaries and for the residual field, so-called the pinned field in [12], which vanishes at the boundaries, its KLT is the sine transform. The detailed PST process as well as the proposed watermark embedding method based on this transform are found in the next section.

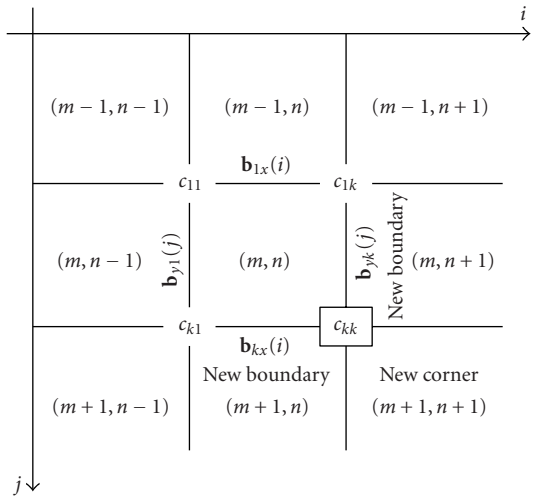


Figure 2: The dual-field decomposition in PST for a typical block.

3. WATERMARK EMBEDDING

The watermark embedding process is described in Figure 1. The details are described as follows. The original image $\mathbf{X}$ is partitioned into non-overlapping blocks of size $k \times k$ as shown in Figure 2. Consider a typical block $\mathbf{X}_{m,n}$, where m and n are the coordinate numbers of this block, we define its corner response as

$$\mathbf{c}_{m,n} = (c_{11}, c_{1k}, c_{k1}, c_{kk}) \tag{3}$$

and its boundary response as

$$\mathbf{b}_{m,n} = (\mathbf{b}_{1x}, \mathbf{b}_{kx}, \mathbf{b}_{y1}, \mathbf{b}_{yk}) \tag{4}$$

as illustrated in Figure 2. The corner response is obtained using the corner function

$$\mathbf{c}_{m,n} = \mathbb{C}[\mathbf{X}_{u,v} : m-1 \le u \le m+1,\ n-1 \le v \le n+1]. \tag{5}$$

More specifically, the corner function is defined as follows:

$$\begin{aligned}
c_{11} &= \frac{\mathbf{X}_{m,n}(1,1)+\mathbf{X}_{m-1,n-1}(k,k)+\mathbf{X}_{m-1,n}(k,1)+\mathbf{X}_{m,n-1}(1,k)}{4},\\
c_{1k} &= \frac{\mathbf{X}_{m,n}(1,k)+\mathbf{X}_{m-1,n}(k,k)+\mathbf{X}_{m-1,n+1}(k,1)+\mathbf{X}_{m,n+1}(1,1)}{4},\\
c_{k1} &= \frac{\mathbf{X}_{m,n}(k,k)+\mathbf{X}_{m,n-1}(k,k)+\mathbf{X}_{m+1,n-1}(1,k)+\mathbf{X}_{m+1,n}(1,1)}{4},\\
c_{kk} &= \frac{\mathbf{X}_{m,n}(k,k)+\mathbf{X}_{m,n+1}(k,1)+\mathbf{X}_{m+1,n}(1,k)+\mathbf{X}_{m+1,n+1}(1,1)}{4};
\end{aligned} \tag{6}$$

and the boundary response is defined by the boundary function

$$\mathbf{b}_{m,n} = \mathbb{B}[\mathbf{X}_{u,v} : m-1 \le u \le m+1,\ n-1 \le v \le n+1] \tag{7}$$

which is further defined as follows:

$$\begin{aligned}
\mathbf{b}_{1x}(i) &= \frac{\mathbf{X}_{m,n}(1,i)+\mathbf{X}_{m-1,n}(k,i)}{2},\\
\mathbf{b}_{kx}(i) &= \frac{\mathbf{X}_{m,n}(k,i)+\mathbf{X}_{m+1,n}(1,i)}{2},\\
\mathbf{b}_{y1}(j) &= \frac{\mathbf{X}_{m,n}(j,1)+\mathbf{X}_{m,n-1}(j,k)}{2},\\
\mathbf{b}_{yk}(j) &= \frac{\mathbf{X}_{m,n}(j,k)+\mathbf{X}_{m,n+1}(j,1)}{2}.
\end{aligned} \tag{8}$$

As we can see from (5)–(8), the processing of one block should involve all the blocks surrounding it, and we can observe in Figure 2 that in a sequential processing of blocks, only one new corner c_{kk} and two new boundaries $\mathbf{b}_{kx}$ and $\mathbf{b}_{yk}$ are required to be computed for a new input block.

The boundary field of $\mathbf{X}_{m,n}$ is achieved by the pinning function [12]

$$\mathbf{X}^b_{m,n} = \mathbb{P}[\mathbf{c}_{m,n}, \mathbf{b}_{m,n}]. \tag{9}$$

Corresponding to the above general form, the specific form of the pinning function is defined as follows:

$$\begin{aligned}
\mathbf{X}^b_{m,n}(i,j) = {} & \mathbf{X}_{m,n}(1,1) + (c_{1k}-c_{11})\frac{(i-1/2)}{k}\\
& + (c_{k1}-c_{11})\frac{(j-1/2)}{k}\\
& + (c_{11}+c_{kk}-c_{k1}-c_{1k})\frac{(i-1/2)(j-1/2)}{k^2}\\
& + \mathbf{g}_x(i) + (\mathbf{h}_x(i)-\mathbf{g}_x(i))\frac{j-1/2}{k}\\
& + \mathbf{g}_y(j) + (\mathbf{h}_y(j)-\mathbf{g}_y(j))\frac{i-1/2}{k},
\end{aligned} \tag{10}$$

where

$$\begin{aligned}
\mathbf{g}_x(i) &= \mathbf{b}_{kx}(i) - \left(c_{k1} + \frac{c_{kk}-c_{k1}}{k}\left(i-\frac{1}{2}\right)\right),\\
\mathbf{h}_x(i) &= \mathbf{b}_{1x}(i) - \left(c_{11} + \frac{c_{1k}-c_{11}}{k}\left(i-\frac{1}{2}\right)\right),\\
\mathbf{g}_y(j) &= \mathbf{b}_{yk}(j) - \left(c_{1k} + \frac{c_{kk}-c_{1k}}{k}\left(j-\frac{1}{2}\right)\right),\\
\mathbf{h}_y(j) &= \mathbf{b}_{y1}(j) - \left(c_{11} + \frac{c_{k1}-c_{11}}{k}\left(j-\frac{1}{2}\right)\right)
\end{aligned} \tag{11}$$

are the pinned boundaries. The pinned field $\mathbf{X}^{p}_{m,n}$ is then given by

$$\mathbf{X}^{p}_{m,n} = \mathbf{X}_{m,n} - \mathbf{X}^{b}_{m,n}. \tag{12}$$

Next, we perform a sine transform to this pinned field block as follows:

$$\mathbf{X}^{p(s)}_{m,n} = \mathbf{S}_k \mathbf{X}^{p}_{m,n} \mathbf{S}^{T}_{k}, \tag{13}$$

where $\mathbf{S}_k$ is the sine transform matrix of order k which is defined as [15]

$$\mathbf{S}_k(i,j) = \sqrt{\frac{2}{k+1}} \sin \frac{\pi(i+1)(j+1)}{k+1}, \tag{14}$$

where $0 \le i, j \le k-1$.

We use a pseudorandom binary sequence as the watermark for image authentication. The length of the sequence L and its initial state number is contained as a part of the secret key file $\mathcal{K}$. The watermark embedding process proceeds by embedding the Pseudorandom sequence into each sine transformed pinned-field block.

Consider a certain transformed block $\mathbf{X}^{p(s)}_{m,n}$; we denote it as

$$\mathbf{X}^{p(s)}_{m,n} = \left\{ x^{p(s)}_{m,n}[t] \right\} \tag{15}$$

by viewing it column by column and with $t \in \mathcal{T} = \{1, 2, \ldots, k^2\}$. The watermark signal intended to be embedded into this block is marked as

$$W_{m,n} = \{ w_{m,n}[l] \} \tag{16}$$

with $l \in \mathcal{L} = \{1, 2, \ldots, L\}$ and $w_{m,n}[l] \in \{0, 1\}$.

In the middle-to-high frequency bands of $\mathbf{X}^{p(s)}_{m,n}$, we select, according to the length of the watermark sequence L, coefficients for watermarking modulation. Suppose the labelling set of these selected coefficients is denoted as $\mathcal{S} = \{t_1, t_2, \ldots, t_L\}$; the watermarking function is then given by

$$\mathbf{Y}^{p(s)}_{m,n} = \mathbb{F}\left[\mathbf{X}^{p(s)}_{m,n}, W_{m,n}, \mathcal{K} \right], \tag{17}$$

where

$$\mathbf{Y}^{p(s)}_{m,n} = \left\{ y^{p(s)}_{m,n}[t] \right\}, \quad t \in \mathcal{T} \tag{18}$$

is the block of watermarked sine transform coefficients. More specifically, the watermarking function $\mathbb{F}[\cdot]$ is defined as in Algorithm 1.

If $t \in \mathcal{S}$, *then*
  if $w_{m,n}[l_t] = 1$, *then*
    if $x^{p(s)}_{m,n}[t] > \lambda$, *then*
      $y^{p(s)}_{m,n}[t] = x^{p(s)}_{m,n}[t]$
    else
      $y^{p(s)}_{m,n}[t] = \alpha_1$
    end if
  else if $w_{m,n}[l_t] = 0$, *then*
    if $x^{p(s)}_{m,n}[t] < -\lambda$, *then*
      $y^{p(s)}_{m,n}[t] = x^{p(s)}_{m,n}[t]$
    else
      $y^{p(s)}_{m,n}[t] = \alpha_2$
    end if
  end if
else if $t \notin \mathcal{S}$, *then*
  $y^{p(s)}_{m,n}[t] = x^{p(s)}_{m,n}[t]$
End if

Algorithm 1

The variables involved in the problem are the following:

(i) $x^{p(s)}_{m,n}[t]$ is the original coefficient;
(ii) $w_{m,n}[l_t]$ is the watermark to be embedded into $x^{p(s)}_{m,n}[t]$;
(iii) $y^{p(s)}_{m,n}[t]$ is the corresponding watermarked coefficient;
(iv) λ is a sufficiently large threshold of positive value. It can be determined by users; its value will affect the tradeoff between the perceptual quality of the watermarked image and the probability of detection of the watermarking scheme;
(v) α_1 and α_2 are floating point values chosen randomly from $[\lambda/2, \lambda]$ and $[-\lambda, -\lambda/2]$, respectively.

The watermarked pinned field block is obtained by the inverse 2D sine transform

$$\mathbf{Y}^{p}_{m,n} = \mathbf{S}^{T}_{k} \mathbf{Y}^{p(s)}_{m,n} \mathbf{S}_k \tag{19}$$

and a watermarked block is therefore achieved by

$$\mathbf{Y}_{m,n} = \mathbf{Y}^{p}_{m,n} + \mathbf{X}^{b}_{m,n}. \tag{20}$$

After processing all the blocks, the watermarked image is the union of all the watermarked blocks:

$$\mathbf{Y} = \bigcup_{m=1}^{M} \bigcup_{n=1}^{N} \mathbf{Y}_{m,n}, \tag{21}$$

where $M \times N$ is the total number of blocks.

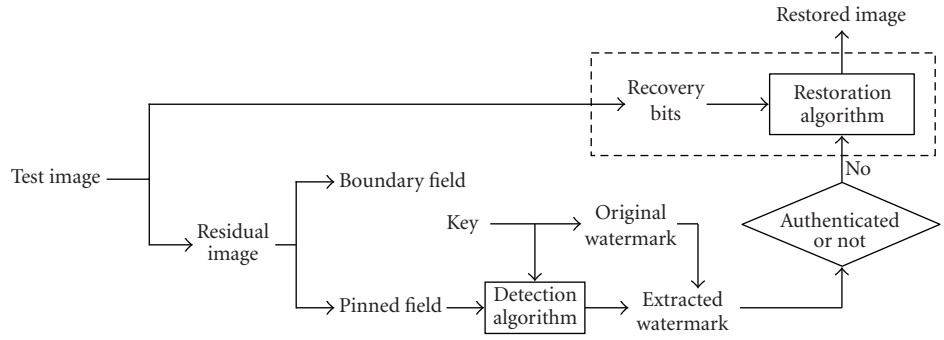


Figure 3: Watermark detection and image authentication process; the parts in the dashed window are optional for host image restoration.

While $t \in \mathcal{S}$ *do*
  if $\hat{y}^{p(s)}_{m,n}[t] \geq 0$, *then*
    $\hat{w}_{m,n}[l_t] = 1$
  else
    $\hat{w}_{m,n}[l_t] = -1$
  End if
End while

Algorithm 2

4. WATERMARK DETECTION, IMAGE AUTHENTICATION AND RESTORATION

The watermark detection and image authentication process is illustrated in Figure 3. The detection system receives as input a watermarked and possibly tampered image $\hat{\mathbf{Y}}$. Similar to the watermarking process, a decomposition is performed on $\hat{\mathbf{Y}}$ by (3)–(12), and then we obtain the sine transform coefficients of its pinned field by (13).

Consider the sine transform components matrix of a certain watermarked pinned filed block:

$$\hat{\mathbf{Y}}^{p(s)}_{m,n} = \left\{ \hat{y}^{p(s)}_{m,n}[t] \right\} \tag{22}$$

by viewing it column by column and with $t \in \mathcal{T} = \{1, 2, \ldots, k^2\}$. The retrieved and possibly corrupted watermark $\hat{W}_{m,n}$ is decided based on the watermark detection function

$$\hat{W}_{m,n} = \mathbb{G}\left[\hat{\mathbf{Y}}^{p(s)}_{m,n}, \mathcal{K} \right]. \tag{23}$$

More specifically, $\mathbb{G}[\cdot]$ is given by Algorithm 2.

$\hat{w}_{m,n}[l_t]$ denotes the watermark bit retrieved from $\hat{y}^{p(s)}_{m,n}[t]$, and $\mathcal{S}$ has the same meaning as in Section 3, which is achieved by the secret key file $\mathcal{K}$.

The original watermark signal $W_{m,n}$ is also generated using the initial state number in the $\mathcal{K}$, and this binary sequence with elements $\{0, 1\}$ is mapped into a corresponding bipolar sequence with elements $\{-1, 1\}$. The watermark bits are compared via the normalized cross correlation function [16]:

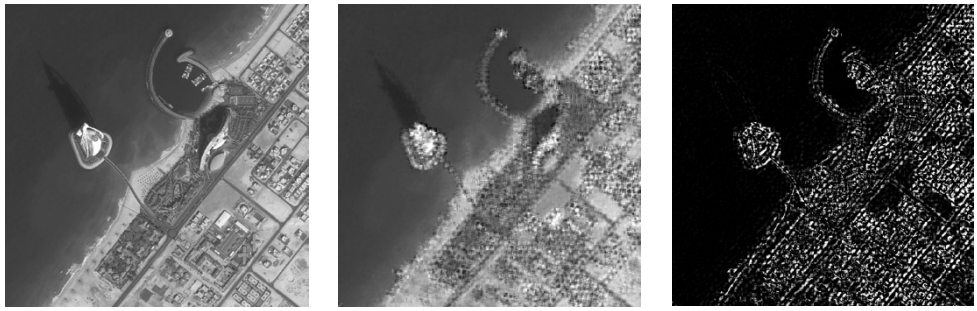
$$\rho = \frac{\sum_{l=0}^{L} \hat{w}_{m,n}[l] w_{m,n}[l]}{\left[\sum_{l=0}^{L} \left(\hat{w}_{m,n}[l] \right)^2 \right]^{1/2} \left[\sum_{l=0}^{L} \left(w_{m,n}[l] \right)^2 \right]^{1/2}}, \tag{24}$$

where $\rho \in [-1, 1]$.

The integrity of the block $\hat{\mathbf{Y}}_{m,n}$ is evaluated according to the value of ρ. If no tampering ever occurred to this block, $\rho \to 1$; on the other hand, ρ will decrease due to different tampering of $\hat{\mathbf{Y}}_{m,n}$. If the content of the block has been changed, that is, the block has been replaced, due to properties of the normalized cross correlation function, ρ will be extremely low.

Assume γ is a properly set threshold; the block is considered to be maliciously tampered with if $\rho < \gamma$. The threshold is determined mathematically or experimentally so as to maximize the probability of detection subject to a given probability of false alarm. In our current simulations, γ is experimentally set to tolerate unavoidable nonmalicious modifications in some practical applications, such as JPEG compression and noise addition, while maintaining the sensitivity of the authentication process to malicious modification on the content of the watermarked images.

If some parts of the watermarked image are detected to be removed or destroyed, these modified regions can be roughly recovered using the method of self-embedding [5]. To facilitate a restoration process, the watermarking embedding and detection processes in Sections 3 and 4 are modified slightly as shown in the dash windows in Figures 1 and 3. In our scheme, the down-sampled image is obtained by compressing the two fields of the original image separately through a sine transform coder as described in [12]. As mentioned in Section 3, for the pinned field, the sine transform coder is equivalent to a fast KLT coder, which results in optimal coding. Another significant advantage of the PST coder over the DCT technique in [5] is that it suppresses significantly the block effect appearing in the recovered image when the compression rate is high by retaining the continuity between blocks [12].



(a) (b) (c)

Figure 4: The dual-field decomposition in the PST of the Dubai image: (a) the original image, (b) the boundary field, and (c) the pinned field.

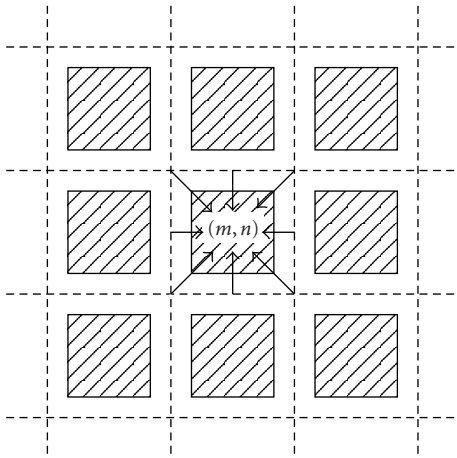


Figure 5: The interblock relationship in the PST.

5. DUAL-FIELD DECOMPOSITION AND SELECTIVE CONTENT AUTHENTICATION

The semifragile watermarking seeks a selective authentication on the content of images. Our scheme aims at protecting the primary textures, such as edges, of the images. To this end, the watermark should not survive the authentication process if such textures are tampered or damaged. The results of the PST dual-field decomposition of the 512×512 *Dubai* image using (3)–(12) are shown in Figure 4. We find that the boundary field is only a blurred version of the original image, while the pinned field is a good characterization of edges, which largely reflects the texture information in the original image. Thus the watermark can be embedded into the pinned field as an indicator of the authenticity of the watermarked image. Moreover, since most common image manipulations tend to preserve such primary features of images, this embedding method ensures that the watermark does not suffer significantly from such legitimate manipulations.

6. INTERBLOCK RELATIONSHIP AND COUNTERFEITING ATTACKS

The most important malicious attacks on existing fragile watermarking schemes are the "cutting and pasting" attacks. The well-known VQ counterfeiting attack proposed by Holliman and Memon [11] is one of such attacks, which thwarts many existing blockwise fragile watermarking methods. In this section, we briefly review the VQ attack by Holliman and Memon and then explain why our scheme can survive the VQ attack.

The success of the VQ attack is based on the assumption that the attacker has a partial knowledge of the possible watermark patterns and it is not restrictive in public applications. The attack starts by collecting a large number of watermarked images, and constructing the codebooks by categorizing all the blocks in those images so that the blocks in the same class correspond to the same watermark pattern. Suppose that the attacker has an unmarked image $\mathbf{Z}$ and intends to counterfeit from it an approximate image $\mathbf{Z}'$ which can pass the authentication system. He examines every block of $\mathbf{Z}$, say, $\mathbf{Z}_{p,q}$, and identifies it as a member of a certain class according to the specific watermarking technique. He then replaces $\mathbf{Z}_{p,q}$ with a watermarked block in that class that minimizes the difference between this block and $\mathbf{Z}_{p,q}$. As thus the attacker achieves his goal without being detected by the authentication system.

In our scheme, we exploit the intrinsic interblock dependence in the PST to detect the above counterfeiting attacks. The "PST style" encoding in (3)–(12) introduces an interblock relationship to the PST images as shown in Figure 5. Therefore, the watermarking of any particular block also depends on its location in the image instead of depending only on its own content. Thus, simple VQ counterfeiting attack can be exposed by this encoding style since the counterfeit of one block affects all the blocks around it; and the construction of codebooks would be very difficult for the reason that the identification of one block should take all the blocks around it into account.

(a) (b) (c)

Figure 6: The original images: (a) Couple, (b) Tank, and (c) Pyramids.

(a) (b) (c)

Figure 7: The watermarked images with recovery bits.

7. EXPERIMENTAL RESULTS

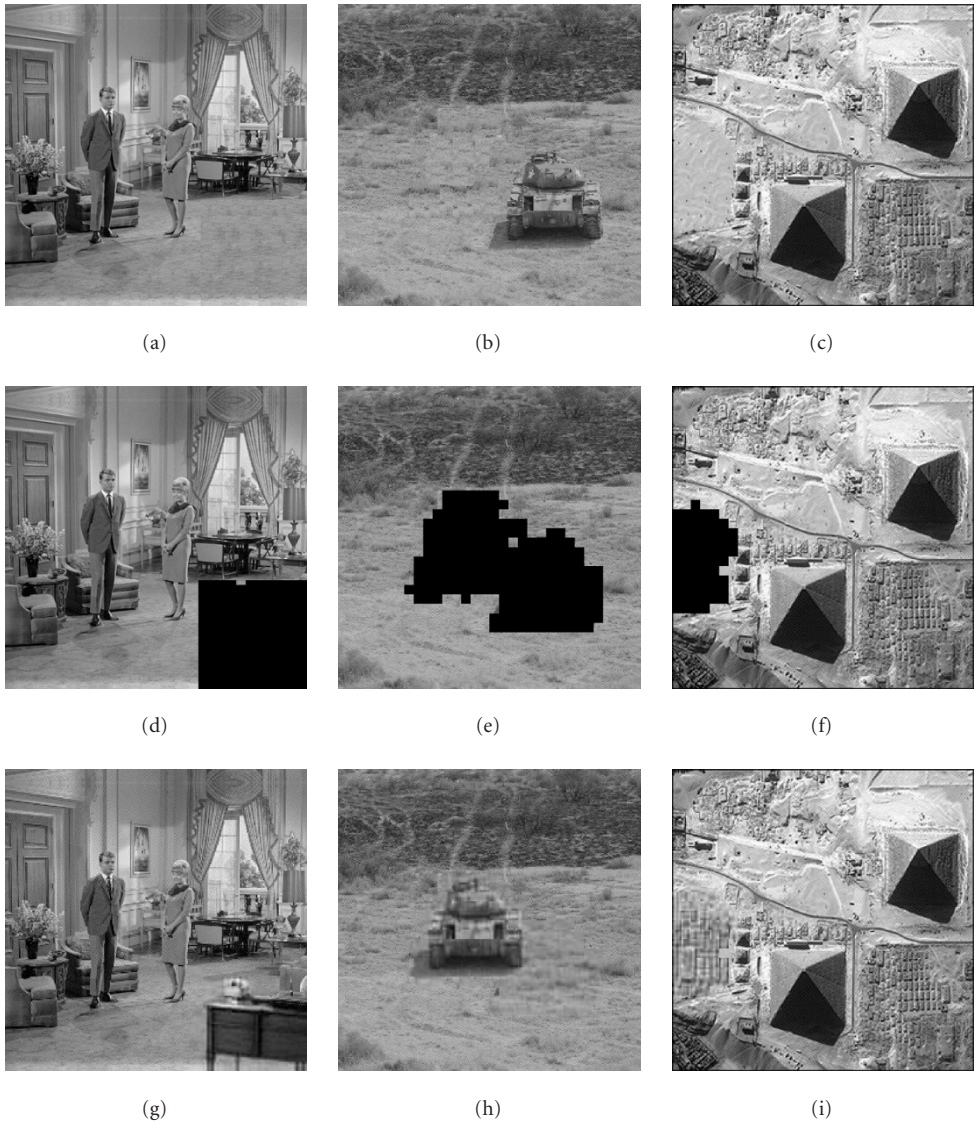
Three 512×512 gray-scale images with different contents and textures were used to test our authentication algorithm. The block size in our experiments was 8×8. The original images are shown in Figure 6. The images shown in Figures 6a and 6b are simple natural images, while Figure 6c is a satellite image with complex texture and fine details. Figure 7 displays the respective watermarked image. We can see that the watermarked images look identical to the original images, with PSNR greater than 33 dB.

We modified the content of the watermarked images in a similar way to the cutting and pasting attacks: all the modifications were performed by cutting and pasting blocks in the same or similar watermarked images. The modification results are shown in Figures 8a–8c. The modifications made to the respective images are as follows: the table in the bottom right corner was removed from the Couple image; the tank was shifted in the Tank image; and in the Pyramids image, some geographical textures were modified. As illustrated in Figures 8d–8f, the modified areas were accurately detected and identified. The approximately recovered images are also presented in Figure 8, which are shown to be visually acceptable. We define the probability of tamper detection P_{TD} of the authentication scheme as

$$P_{\mathrm{TD}} = \frac{\mathrm{NUM}_{\mathrm{detected}}}{\mathrm{NUM}_{\mathrm{modified}}}, \tag{25}$$

where $\mathrm{NUM}_{\mathrm{modified}}$ is the number of actually modified blocks, and $\mathrm{NUM}_{\mathrm{detected}}$ is the number of correctly detected blocks. In our experiments, P_{TD} without nonmalicious attacks was always higher than 98%.

We also tested the insensitivity of our algorithm to compression. As shown in Figure 9, before compression, the output ρ of the watermark detection system sharply peaked at 1; after compression, the values of ρ decreased as shown in the same figure. To illustrate the advantage of PST watermarking, we compare the performance of PST watermarking with that of DCT watermarking. In the DCT watermarking, the same watermark embedding method was used and the same middle frequency-band coefficients were selected as those in the PST watermarking. The comparison was based on the same PSNR values of the watermarked images and the results were obtained through averaging the outcomes of the three test images. We found that after the



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figure 8: Sample results of the proposed watermarking scheme: (a)–(c) modified images, (d)–(f) authentication outputs, and (g)–(i) restoration outputs.

compression, the drop in the detector output ρ for the PST watermarking was smaller than that of the DCT watermarking. This indicates that the PST watermarking is less sensitive to JPEG compression than the DCT watermarking, which makes it a better candidate for semifragile watermarking. Given a certain value of the threshold γ, the probability of detection P_D is shown as the shaded area in Figure 9. It is apparent from this figure that the P_D of the PST scheme is larger than that of DCT. The collective comparison results with $\gamma = 0.1$ and varying compression quality factor (QF) values are reported in Figure 10. The higher values of P_D indicated the better detection performance of PST over DCT. Even when the images were in very poor quality as shown in Figure 11, the P_D of our scheme was still higher than 95%.

The performance of our algorithm against JPEG compression and additive noise from Stirmark 4[1] was also tested. After content modification, the watermarked image in Figure 8a was JPEG compressed with a QF of 90% and the watermarked image in Figure 8c is added with an additive white Gaussian noise of zero mean and a variance of $\sigma^2 = 5$, as shown in Figures 12a and 12b, respectively. As the recovery bits were simply inserted into the pixels' LSBs, the recovery results are no longer correct. However, such manipulations only have minimum effect on the authentication process. As indicated in Figures 12c and 12d, the modified area still can be correctly identified.

[1] www.cl.cam.ac.uk/fapp2/watermarking/stirmark.

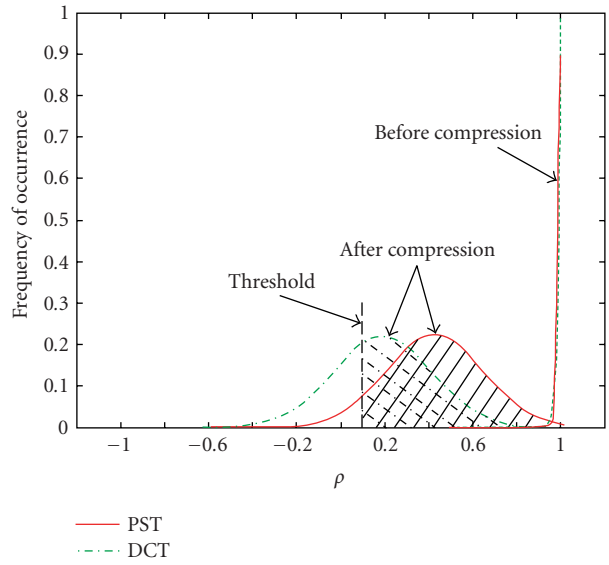


Figure 9: The distribution of the watermark detection outputs before and after JPEG compression (QF = 40).

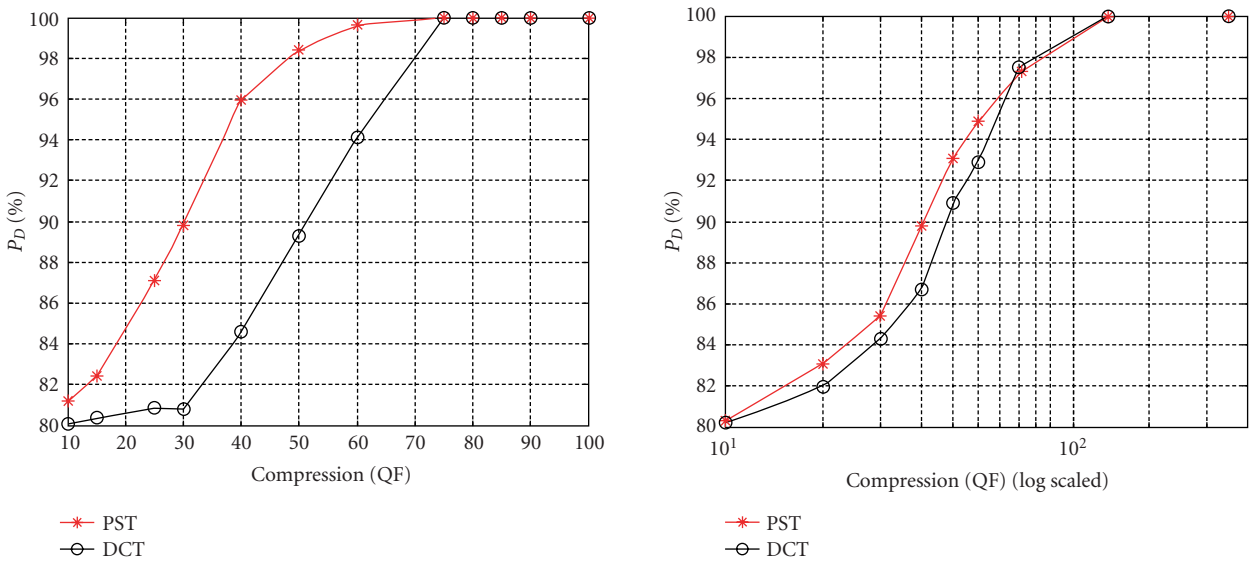


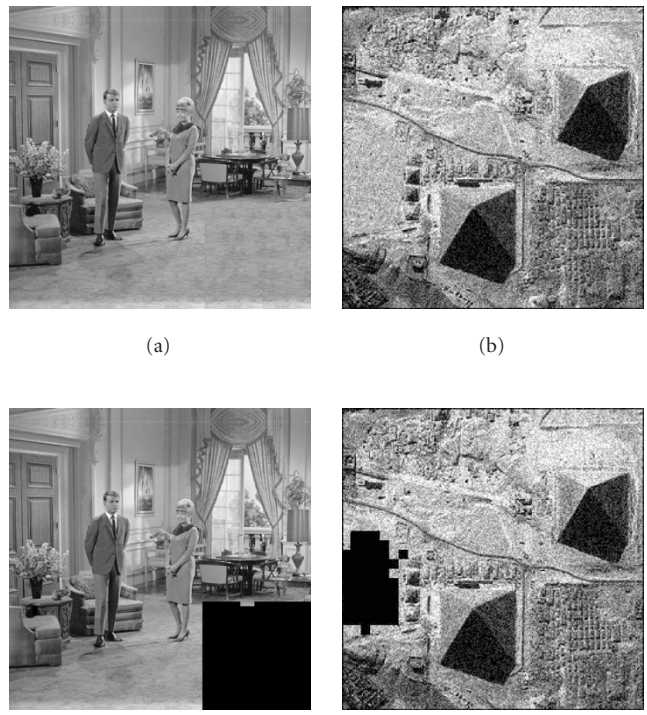
Figure 10: Comparisons between PST watermarking and conventional DCT watermarking: the probability of detection after (a) JPEG compression and (b) wavelet compression.

8. CONCLUSION AND FUTURE WORK

In this paper, we investigated the problem of the selective content authentication of digital images through a novel semifragile watermarking using the pinned sine transform (PST). The watermark is embedded into the pinned field of PST, which contains the texture information of the original image. This important property of the pinned field provides the scheme with special sensitivity to any texture alteration of the watermarked image. The effectiveness of the new method has been demonstrated by using natural scene images and satellite images. In the authentication process, the probability of detection was higher than 98%. The scheme was very robust to cutting and pasting counterfeiting attacks. It was also able to tolerate some common image processing manipulations; the probability of detection after JPEG compression

(a) (b)

Figure 11: Attacked images. (a) Watermarked Couple image after JPEG compression (QF= 40). (b) Watermarked Couple image after wavelet compression (QF= 60).



(a) (b)

(c) (d)

Figure 12: Sample authentication results after JPEG compression and additive noise from Stirmark 4. (a) Watermarked and modified Couple image after JPEG compression (QF= 90). (b) Watermarked and modified Pyramids image with additive noise ($\sigma^2 = 5$). (c) Authentication result of (a). (d) Authentication result of (b).

and wavelet compression is higher than that of equivalent DCT scheme. In future work, we are interested in developing image authentication methods incorporating restoration that can survive various nonmalicious manipulations.

REFERENCES

[1] I. J. Cox, M. L. Miller, and J. A. Bloom, *Digital Watermarking*, Morgan Kauffman Publishers, San Francisco, Calif, USA, 2001.

[2] M. M. Yeung and F. Mintzer, "An invisible watermarking technique for image verification," in *Proc. IEEE International Conference on Image Processing (ICIP '97)*, vol. 2, pp. 680–683, Santa Barbara, Calif, USA, October 1997.

[3] P. W. Wong, "A watermark for image integrity and ownership verification," in *Proc. IS & T's Image Processing, Image Quality, Image Capture, Systems Conference (PICS '98)*, pp. 374–379, Portland, Ore, USA, May 1998.

[4] P. W. Wong, "A public key watermark for image verification and authentication," in *Proc. IEEE International Conference on*

Image Processing (ICIP '98), vol. 1, pp. 455–459, Chicago, Ill, USA, October 1998.

[5] J. Fridrich and M. Goljan, "Images with self-correcting capabilities," in *Proc. IEEE International Conference on Image Processing (ICIP '99)*, vol. 3, pp. 792–796, Kobe, Japan, October 1999.

[6] C.-Y. Lin and S.-F. Chang, "Semifragile watermarking for authenticating JPEG visual content," in *Security and Watermarking of Multimedia Contents II*, vol. 3971 of *Proceedings of SPIE*, pp. 140–151, San Jose, Calif, USA, January 2000.

[7] D. Kundur and D. Hatzinakos, "Digital watermarking for telltale tamper proofing and authentication," *Proceedings of the IEEE*, vol. 87, no. 7, pp. 1167–1180, 1999.

[8] C.-S. Lu and H.-Y. M. Liao, "Multipurpose watermarking for image authentication and protection," *IEEE Trans. Image Processing*, vol. 10, no. 10, pp. 1579–1592, 2001.

[9] L. Me and G. R. Arce, "A class of authentication digital watermarks for secure multimedia communication," *IEEE Trans. Image Processing*, vol. 10, no. 11, pp. 1754–1764, 2001.

[10] M. U. Celik, G. Sharma, E. Saber, and A. M. Tekalp, "Hierarchical watermarking for secure image authentication with localization," *IEEE Trans. Image Processing*, vol. 11, no. 6, pp. 585–595, 2002.

[11] M. Holliman and N. Memon, "Counterfeiting attacks on oblivious block-wise independent invisible watermarking schemes," *IEEE Trans. Image Processing*, vol. 9, no. 3, pp. 432–441, 2000.

[12] A. Z. Meiri and E. Yudilevich, "A pinned sine transform image coder," *IEEE Trans. Communications*, vol. 29, no. 12, pp. 1728–1735, 1981.

[13] A. K. Jain, "Some new techniques in image processing," in *Proc. ONR Symposium on Current Problems in Image Science*, O. Wilde and E. Barrett, Eds., pp. 201–223, Monterey, Calif, USA, November 1976.

[14] A. Z. Meiri, "The pinned Karhunen-Loeve transform of a two dimensional Gauss-Markov field," in *Proc. SPIE Conference Image Processing*, San Diego, Calif, USA, 1976.

[15] A. K. Jain, *Fundamentals of Digital Image Processing*, Prentice-Hall, Englewood Cliffs, NJ, USA, 1989.

[16] W. K. Pratt, *Digital Image Processing*, John Wiley & Sons, New York, NY, USA, 2nd edition, 1991.

Anthony T. S. Ho is currently an Associate Professor in the Division of Information Engineering, School of Electrical and Electronic Engineering, Nanyang Technological University (NTU). He is also the Program Director for Digital Watermarking, Centre for Information Security at NTU. He co-founded DataMark Technologies (DMT) in 1998 which specializes in digital watermarking and steganography. He obtained his B.S. degree (Honors) in physical electronics from the University of Northumbria, UK, in 1979, his M.S. degree in applied optics from Imperial College of Science, Technology and Medicine, University of London, in 1980, and his Ph.D. degree in digital image processing from King's College, University of London, in 1983. He is a Fellow of the Institution of Electrical Engineers (FIEE), a Chartered Electrical Engineer (C.Eng.), and a Senior Member of the Institute of Electrical and Electronic Engineers (SMIEEE). He also serves as a Director on the Board of DMT and provides consultancy to the company.

Xunzhan Zhu was born in 1980 in Zhejiang, China. She received her B.S. degree in 2002 in communication engineering from Beijing University of Posts and Telecommunications, China. She is now a Ph.D. candidate at the School of Electrical and Electronic Engineering at Nanyang Technological University (NTU), Singapore. Her current research area is digital image watermarking.

Yong Liang Guan received his B.Eng. and Ph.D. degrees from the National University of Singapore and Imperial College of Science, Technology and Medicine, University of London, respectively. He is currently an Assistant Professor at the School of Electrical and Electronic Engineering, Nanyang Technological University (NTU), Singapore. He is also the Program Director of the Wireless Network Research Group in the Positioning and Wireless Technology Center (PWTC), and the Deputy Director of the Center for Information Security, NTU. His research interests include digital multimedia watermarking, advanced modulation and coding, and broadband channel modeling.