Wavelet-based Fragile Watermarking Scheme for Image Authentication

Chang-Tsun Li and Huayin Si

Department of Computer Science
University of Warwick
Coventry, CV4 7AL, UK
{ctli, waynesi@dcs.warwick.ac.uk}

August 2006

ABSTRACT

In this work, we propose a novel fragile watermarking scheme in wavelet transform domain, which is sensitive to all kinds of manipulations and has the ability to localize the tampered regions. To achieve high transparency (i.e., low embedding distortion) while providing protection to all coefficients, the embedder involves all the coefficients within a hierarchical neighborhood of each sparsely selected watermarkable coefficient during the watermark embedding process. The way the non-watermarkable coefficients are involved in the embedding process is content-dependent and non-deterministic, which allows the proposed scheme to put up resistance to the so-called vector quantization attack, Holliman-Memon attack, collage attack and transplantation attack.

Keywords: fragile watermarking, digital watermarking, authentication, content verification, multimedia security.

1. INTRODUCTION

In recent years, the rapid expansion of the interconnected networks and the never-ending development of digital technologies have facilitated instant multimedia transmission and the creation of large-scale digital image databases. The advantages of digitized images are that images can be easily manipulated and reproduced without significant loss of quality. However, these also imply that images can be modified easily and imperceptibly with malicious intentions. Techniques securing information flowing on the networks are therefore essential for protecting intellectual properties and fostering the development of electronic commerce. To meet the security needs, researchers have been actively investigating techniques of digital watermarking in the last decade [1-7]

To achieve the goals of copyright protection and ownership identification, *robust* watermarking schemes have been developed. They are expected to survive various kinds of manipulation to a reasonable extent provided that the altered media is still acceptable in terms of visual quality, or valuable in terms of commercial significance [1, 5, 8-10]. Another application of digital watermarking is multimedia authentication and content integrity verification. This kind of schemes focuses on the capability of detecting forgeries. Therefore, this type of watermarks [2, 6, 7, 11-14] is usually *fragile* and is expected to be sensitive to attacks. The proposed work falls in this category.

From the attacker's point of view, the goal of attacking the fragile watermarking schemes is to alter the image while keeping the watermark intact in the meantime. Cut-and-paste is probably one of the most likely attacks watermarking schemes may face. It is about cutting one area from the same image or other images and pasting it somewhere else in the image, intending to change the content or semantics of the image.

Although naive, the cut-and-paste attack is most likely to succeed if the watermarking scheme uses the same content-independent watermark to authenticate a large set of image, e.g., an image database. One common way of countering this attack is to partition the image into blocks and use the encrypted output of a hash function, which takes the image block as input, as the watermark. However, according to *birthday paradox* [15] if the attacker has $2^{l/2}$ watermarked blocks available, where l is the length of the hash output, then the possibility of finding two blocks with the same hash output / watermark is 0.5. Based on birthday paradox, a counterfeit can be forged by combining blocks taken from a large image database without knowing the secret key. This form of attack is referred to as vector quantization attack [16], the Holliman-Memon attack [17], or collage attack [18].

As pointed out in [2, 7, 11, 17-19], the essential requirement of countering the afore-mentioned attacks is to establish block-wise dependence or to introduce contextual information so that watermark embedding involves not only the block / pixel itself but also other blocks / pixels within a neighborhood. With the involvement of the contextual information, the vector quantization attack cannot succeed because placing watermarked blocks in the wrong context will not pass the authentication. But as Barreto et al. have observed, if the contextual information is calculated, not in a random, but in a deterministic manner, the scheme is still vulnerable to the transplantation attack [11], which is another form of malicious operation of collecting blocks to create a counterfeit. For example, let $f'_A \rightarrow f'_B$ denote that the signature of block f'_B is generated based on the information about f'_A. For two images, f' and f'', with block f'_A, f'_B, f'_C identical to f''_A, f''_B, and f'_C, respectively, but f'_X not identical to f''_X, if the following dependence relationships

$$\ldots \rightarrow f'_A \rightarrow f'_X \rightarrow f'_B \rightarrow f'_C \rightarrow \ldots$$

$$\ldots \rightarrow f''_A \rightarrow f''_X \rightarrow f''_B \rightarrow f''_C \rightarrow \ldots$$

exist, then the pairs (f'_X, f'_B) and (f''_X, f''_B) can be swapped without being detected by schemes exploiting deterministic dependence. This type of transplantation attack can still be successful even if the number of dependencies is increased. For example, let $f_A \leftrightarrow f_B$ denote that the signature of each block depends on the information about the other. Now if the following dependence relationships

$$\cdots \leftrightarrow f'_A \leftrightarrow f'_B \leftrightarrow f'_X \leftrightarrow f'_C \leftrightarrow f'_D \leftrightarrow \cdots$$

$$\cdots \leftrightarrow f''_A \leftrightarrow f''_B \leftrightarrow f''_X \leftrightarrow f''_C \leftrightarrow f''_D \leftrightarrow \cdots,$$

exist, then triplets (f'_B ,f'_X , f'_C) and (f''_B ,f''_X , f''_C) are interchangeable without being noticed if block f'_D is also identical to f''_D. Therefore, to thwart the transplantation attack, non-deterministic dependence has to be imposed as a key requirement on the watermarking schemes. The reader is referred to [7, 16-18] and [11, 19] for more information about vector quantization attack and transplantation attack, respectively.

Based on the above discussions, we believe non-deterministic contextual dependence information should be taken as one of the key requirement of an effective fragile watermarking scheme.

Generally speaking, fragile watermarking can be classified into spatial-domain and transform-domain approaches. Usually, spatial-domain fragile watermarking schemes [11, 13, 14, 16, 19] watermark all the pixels. However, they are not directly applicable to where transformation and quantization are necessary for compressing the images because each small level of the quantized coefficient value corresponds to one big quantization step. This makes exhaustive embedding a visually intrusive operation. To maintain low embedding distortion, transform-domain watermarking schemes [20-22] tend to

watermark some selected coefficients in the mid-frequency of the host image. However, we observed [2] that many of the schemes [20-22] watermarked only some selected coefficients while leaving most coefficients unprotected in order to minimize the embedding distortion. As a result, a wide security gap is left open to attacks. Our observation suggests that measures of protecting *all* the coefficients without actually watermarking all coefficients and compromising the visual quality of the image are desirable.

In recent years, the standardization process of JPEG 2000 and the trend of shifting from DCT to Discrete Wavelet Transform (DWT) based image compression have prompted the development of some watermarking schemes in wavelet transform domain for the applications of image authentication [20-25]. It is our intention in this work to propose a novel fragile watermarking scheme for authenticating JPEG2000 images.

The rest of this work is organized as follows. Sec. 2 reviews some related works. Sec. 3 describes *what* the proposed scheme does. Analyses are conducted in Sec. 4 to clarify *why* the algorithm does what described in Sec. 3. Simulations are carried out in Sec. 5 to test the proposed scheme. Finally, Sec. 6 concludes this work.

2. RELATED WORKS

In Yuan and Zhang's work [24], a Gaussian mixture statistical model is used to get the distribution parameters of wavelet coefficients. Some coefficients of large value are modified to embed the watermark. Dependence among one watermarkable block and n^2-1 reference blocks are established in the embedding process. However, since the

neighborhood is big, when the image is tampered with, their scheme cannot localize the tampering accurately.

In [25], Paquet et al. applied wavelet-packets decomposition to a target image, and used a key to select the level of details and coefficients in which the watermark is to be embedded. Their scheme established the block-wise dependence using the key. Although they exploited the characteristics of human visual system (HVS) to minimize the distortion, the embedding operation introduced significant distortion in terms of PSNR with an average at 42.38 dB.

The scheme due to Xie and Arce [21] extracted and encrypted the edge of the approximation component after DWT, and then etched it into the approximation. Since, for the sake of robustness, they did not protect the details of the image, leaving a security gap open to the attacker.

In the approach proposed by Winne et al. [22], to minimize the embedding distortion and maintain high localization accuracy, only the coefficients of the high-frequency sub-bands at the finest scale of the luminance component are watermarked. All the other coefficients and components are neither watermarked nor involved during the watermarking process of the embeddable coefficients. Moreover, another limitation of their algorithm is that they do not establish block-wise dependence during the embedding process. Due to the lack of mutual dependence, this scheme is vulnerable to cut-and-paste, vector quantization, and transplantation attacks [11].

Our observation suggests that watermarking schemes capable of overcoming the afore-mentioned problems are desirable. The intention in this work is to propose a fragile watermarking scheme with the following capabilities:

1 Achieve high resolution of tamper localization

2 Maintain low embedding distortion.

3 Protect all the coefficients without actually watermarking them all.

4 Exploit non-deterministic block-wise dependence to thwart the afore-mentioned attacks.

3. PROPOSED SCHEME

The proposed scheme can be incorporated into JPEG2000 pipeline to facilitate authentication. JPEG2000 [26] is a modern standard, comprising six stages shown as solid blocks in Fig. 1, for image compression. The decoder works in the reverse order. Our proposed watermarking scheme, shown as dotted block in Fig. 1, is invoked after quantization (stage 4) and prior to entropy coding (stage 5). Authentication is carried out between the same stages in a reverse manner. In this work, the 'Haar' or 'Daubechies db1' wavelet base is used for DWT, but the algorithm is applicable in conjunction with other wavelet bases. Since the proposed scheme watermarks the quantized coefficient (i.e., it is independent of the quantization table) and the watermarked coefficients will be coded by the ensuing entropy coding stage, therefore it is compatible with JPEG2000 standard regardless what quantization table is used.

The idea is described as follows: Given a DWTed image X with n decomposition level as demonstrated in Fig. 2, a binary sequence A of the same dimension and structure as X is generated with a secret key. Another binary map B of the same dimension and structure as X is created such that all its pixels corresponding to the non-zero-valued

coefficients in X are set to 1 and the others set to 0. But B_{HH1} is modified with a 'projection' operation to be described later. In order to meet the requirement of low embedding distortion, in this work we will watermark some of the selected coefficients in X_{HH1} sub-band only. B_{HH1} serves as a map indicating which coefficients in X_{HH1} sub-band are watermarkable. A binary watermark W is created by performing EXCLUSIVE-OR operation on the binary map A and B. Then for every selected coefficient in $X_{HH1}(i,j)$, a secret sum $S(i,j)$ is calculated by summing up the coefficients from a hierarchical neighborhood $R(i,j)$ in a secret / random manner according to their corresponding watermark bits in W. To embed a watermark bit in a selected coefficient, the selected coefficient is modulated such that the corresponding watermark bit $W(i,j)$ is equal to a binary function, which takes the selected coefficient $X_{HH1}(i,j)$ and the secret sum $S(i,j)$ as inputs . A candidate for the binary function could be the parity of $X_{HH1}(i,j) + S(i,j)$ or some sort of combination of $X_{HH1}(i,j)$ and $S(i,j)$. The watermarking process repeats until all the selected coefficients are marked. To authenticate and verify the watermarked image, the verifier / authenticator performs the same operations as applied on the embedding side to calculate the secret sum and compares the binary function with the watermark generated with the same secret key shared between the embedder and verifier / authenticator.

In order to make the algorithm clearer, before the presentation of the algorithms some symbols are defined and explained as follows.

X: The set of quantized DWT coefficients of the original image. X is a two-dimensional matrix with a structure of bands as shown in Fig. 2 and can be represented as

$$X = \{X_{LLn}, X_{HLn}, X_{LHn}, X_{HHn}, X_{HL(n-1)}, X_{LH(n-1)}, X_{HH(n-1)}, \cdots, X_{HL1}, X_{LH1}, X_{HH1}\}$$

where n is the number of DWT decomposition levels and subscripts LL_l, HL_l, LH_l, HH_l identify the sub-bands at level l, $l \in [1, n]$.

A: A binary sequence with the same dimensions and structure as X generated with a secret key shared between the embedder and verifier / authenticator.

B: A binary sequence with the same dimensions and same structure as X. Each coefficient in B, except the ones in B_{HH1} sub-band, is assigned a value 1 if its corresponding DWT coefficient in the same position of X is not zero; otherwise, 0 is assigned instead. B_{HH1} is modified by projecting all the bands in every level to this band. The procedures of the projection are described as follows:

1) Perform the logical *OR* operation on all of the three sub-bands at each level of B (four in the highest level n including B_{LLn}) to create a sequence of binary configuration denoted as $C = \{C_n, C_{n\text{-}1}, \ldots, C_1\}$ according to Eq. (1) and (2).

$$C_n(i,j) = \begin{cases} 1, & \text{if } B_{HLn}(i,j) + B_{LHn}(i,j) + B_{HHn}(i,j) + B_{LLn}(i,j) = 1 \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

$$C_l(i,j) = \begin{cases} 1, & \text{if } B_{HLn}(i,j) + B_{LHn}(i,j) + B_{HHn}(i,j) = 1 \\ 0, & \text{otherwise} \end{cases} \qquad \forall\ l < n \tag{2}$$

where '+' is the logical OR operator.

2) For each $l \in [2, n]$, project C_l to the C_1 *recursively* according to Eq. (3).

$$C_1(i,j) = \begin{cases} C_l\left(\left\lfloor \frac{i}{2^{l-1}} \right\rfloor, \left\lfloor \frac{j}{2^{l-1}} \right\rfloor\right), & \text{if all the descendents of } C_l\left(\left\lfloor \frac{i}{2^{l-1}} \right\rfloor, \left\lfloor \frac{j}{2^{l-1}} \right\rfloor\right) \text{ on } C_1 \text{are 0} \\ C_1(i,j), & \text{otherwise} \end{cases} \tag{3}$$

Note by *recursive* projection, we mean once level C_l has been projected onto C_1, level C_{l+1} will be projected onto the updated C_1. From the first part of Eq. (3), we know that any coefficient having at least one child coefficient with

value 1 will not be projected. Fig. 3 illustrates the idea of the *recursive* projection of C_2 and C_3 onto C_1. The purpose of the projection will be discussed in Sec. 4.1.

3) Substitute the final C_1 for B_{HH1}.

W: The binary watermark sequence resulted from the Exclusive-OR operation on A and B. W also has the same structure as X. Therefore,

$$W = A \oplus B \tag{4}$$

The reader is reminded that the B_{HH1} sub-band of B has been updated according to Eq (1), (2) and (3).

$R(i, j)$: The hierarchical neighborhood of $X_{HH1}(i, j)$, including not only itself and the conventional 8-neighborhood of (i, j), but also two 3×3 blocks in $HL1$ and $LH1$ bands centered at position (i, j) and the corresponding ancestors of (i, j) in all bands at all higher levels. An example of the hierarchical neighborhood $R(i, j)$ of a coefficient is shown in Fig. 4.

$S(i, j)$: The non-deterministic secret sum of $R(i, j)$. When a coefficient $X_{HH1}(i, j)$ is selected to be watermarked, $S(i, j)$ is calculated as follows,

$$S(i,j) = \sum_{(i',j') \in R(i,j)} (-1)^{W(i',j') \oplus W_{HH1}(i,j)} X'(i',j') \tag{5}$$

, where $X'(i', j')$ is defined as

$$X'(i',j') = \begin{cases} \left\lfloor \dfrac{X(i',j')}{2} \right\rfloor \times 2, & \text{if } X(i',j') \text{ is watermarkable} \\ X(i',j'), & \text{otherwise} \end{cases} \tag{6}$$

where $\lfloor \cdot \rfloor$ is the floor function which rounds its argument towards zero. The importance of $S(i, j)$ and the reason of performing the operation in the upper part

of Eq. (6) will be discuss in Sec. 4.2 and 4.3, respectively.

Parity($S(i,j)$, $X_{HH1}(i,j)$): The function *Parity* returns the parity of the result of concatenating the sign of $S(i, j)$ with '0' representing positive and '1' representing negative, the binary string of the *absolute* value of $S(i, j)$, the *sign* of $X_{HH1}(i, j)$ and the binary string of the *absolut*e value of $X_{HH1}(i, j)$. If the number of 1s is even *Parity* returns 0, otherwise it returns 1. For example, if $S(i, j) = -15$ and $X_{HH1}(i, j) = 2$, the concatenated binary string will be $(1)_2\ (1111)_2\ (0)_2\ (10)_2 = (1\ 1111\ 0\ 10)_2$ and *Parity* returns 0.

3.1 Watermark Embedding Algorithm

Now, the proposed watermark-embedding algorithm can be described as follows:

Step$_e$1. Generate X by performing the n-level discrete wavelet decomposition on the target image and quantizing the DWT coefficients.

Step$_e$2. Generate A with the secret key k shared with the authenticator.

Stepe3. Generate B according to X and modify B_{HH1} by performing the projection operation as described in Eq. (1) - (3).

Step$_e$4. Generate W according to Eq. (4).

Stepe5. Take the coefficients in X_{HH1}, whose counterparts in B_{HH1} have a value of 1, as *watermarkable*. For each watermarkable coefficient $X_{HH1}(i,j)$, establish a neighborhood $R(i,j)$ and calculate the secret sum $S(i, j)$ according to Eq. (5) and (6).

Step$_e$6. Embed the watermark bit into the watermarkable coefficient $X_{HH1}(i, j)$ according to the modulation method described below such that Eq. (7) is satisfied

$$Parity(S(i,j), X_{HH1}(i,j)) = W_{HH1}(i,j) \tag{7}$$

Modulation algorithm

Watermark bit $W_{HH1}(i,j)$ is embedded by enforcing Eq. (7) through modulating $X_{HH1}(i,j)$ according to the following tow cases.

Case 1*:* If Eq. (7) *does* hold before modulation, then

$$X_{HH1}(i,j) = \begin{cases} X_{HH1}(i,j), & \text{if } X_{HH1}(i,j) \neq 0 \\ -1, & \text{if } X_{HH1}(i,j) = 0 \end{cases} \tag{8}$$

Case 2*:* If Eq. (7) does *not* hold before modulation, then

$$X_{HH1}(i,j) = \begin{cases} 1, & \text{if } X_{HH1}(i,j) = -1 \\ -1, & \text{if } X_{HH1}(i,j) = 1 \\ bitxor(X_{HH1}(i,j),\ 1), & \text{otherwise} \end{cases} \tag{9}$$

where $bitxor(X_{HH1}(i,j),\ 1)$ is the Exclusive-OR operation on $X_{HH1}(i,j)$ and constant 1. The *bitxor* operation is to reverse the LSB of $X_{HH1}(i,j)$.

3.2 Watermark Authentication Algorithm

Step$_a$1. Restore the wavelet coefficients X by decoding the received bit stream.

Step$_a$2. Generate A with the same secret key k shared with the embedder.

Step$_a$3. Generate B according to X, and modify B_{HH1} by performing the projection operation as described in Eq. (1) - (3).

Step$_a$4. Generate W according to Eq. (4).

Step$_a$5. Take the coefficients in X_{HH1}, whose counterparts in B_{HH1} have a value of 1, as *watermarked*. For each watermarked coefficient $X_{HH1}(i,j)$, establish neighborhood $R(i,j)$ and calculate the secret sum $S(i,j)$ according to Eq. (5) and

(6).

Step$_a$6. Authenticate the watermarked coefficients by verifying whether Eq. (7) holds or not. If Eq. (7) does not hold, the corresponding position in a binary authentication *map* $D(i,j)$ is set to 255 so as to indicate the occurrence of tampering. Otherwise, the corresponding pixel in the authentication map is set to 0.

The processes of watermark embedding and authenticating are illustrated in Fig. 5.

4. ALGORITHM ANALYSES

4.1 Reducing Embedding Distortion

To minimize embedding distortion, watermarking should take place in as few bands and coefficients as possible. Since human visual system is less sensitive to the information in high frequency bands and diagonal noise patterns [27] and X_{HH1} band contains the coefficients of highest frequencies with diagonally oriented features, in the proposed scheme, only the coefficients in X_{HH1} band, whose counterpart in B_{HH1} have a value equal to 1, are selected for watermarking.

To provide protection to all the unwatermarked coefficients without actually marking them all, association among the watermarkable and unwatermarkable coefficients must be established so that every coefficient will have their 'representative', the watermarkable one, in X_{HH1} band. This is achieved by creating B and then recursively projecting all the bands of B in all levels onto B_{HH1}. The operation of projecting B ensures that any coefficients will have at least one watermarkable descendant or sibling in X_{HH1} band. And

the dependence S established upon the hierarchical neighborhood R ensures that if any members of the neighborhood are attacked, their representative(s) will raise alarm for them. This is because if a coefficient is attacked, at least one of $X(i, j)$, $B(i,j)$ *or* $S(i,j)$ will be changed, making correct watermarks extraction impossible and consequently failing the authentication.

There is no general rule of thumb for deciding the number of decomposition levels. The choice depends on how the user weighs the importance of the contradicting factors of security and resolution of tamper localization. To emphases the importance of security the user could decompose the image into greater number of levels so that each coefficient at the coarsest level can have more watermarkable descendents in the B_{HH1} band. If the emphasis is on resolution of tamper localization, then smaller number of levels should be used.

4.2 Thwarting the Vector Quantization Attack and Transplantation Attack

As pointed out in Sec. 1, non-deterministic contextual dependence is the key requirement for thwarting transplantation attack, vector quantization attack, Holliman-Memon attack and collage attack. We can see from Eq. (5) that this requirement is achieved by involving the watermark bits $W(i,j)$ and $W'(i,j)$, which are created through the combination of the random binary map A and B with Eq. (4). Actually, the security of the whole scheme relies on the secret key, which is unknown to the attacker. According to Equation (5), the secret sum S is a function of W, which, according to equation (4), is in turn a function of A and A is generated with the secret key. Therefore with the secret key and W unavailable to the attacker, he/she cannot attack the DWT coefficients while

preserving S because S is also unknown to him/her.

4.3 Consistency of Secret Sum Used by the Embedder and Authenticator

To facilitate authentication, the *same* secret sum $S(i, j)$ has to be used by the embedder and verifier / authenticator. From Eq. (8) and (9), which formulate the watermark embedding, we can see that a watermarkable coefficient may undergo changes of the following forms:

1) from 0 to –1 according to part two of Eq. (8)
2) from -1 to 1 according to part one of Eq. (9)
3) from 1 to -1 according to part two of Eq. (9)
4) LSB reversed according to part three of Eq. (9)

For the first three forms of change, no matter the watermarkable coefficients are equal to 0, 1, or –1, the operation in the upper part of Eq. (6) returns the same value of 0, i.e. the changes make no difference in the calculation of the secret sum $S(i, j)$ defined in Eq. (5). For the fourth form of change, the operation in the upper part of Eq. (6) excludes the least significant bit (LSB) of a watermarkable coefficient in the calculation of $S(i, j)$. By preventing the changes due to the watermark embedding process from involving in the calculation of the secret sum $S(i, j)$ we can ensure that the embedder and verifier sides are using the same value of $S(i, j)$.

5. SIMULATION RESULTS

This section presents some simulation to verify the capacities of the proposed

watermarking scheme. All the images are of 256×256 pixels, with the intensity of each pixel represented in 8 bits. They are wavelet-decomposed into 3 levels and quantized with the quantization step $Q = 5$.

5.1 Visibility Evaluation

We watermarked four images, namely Lena, Barbara, Cameraman and Mandrill. To evaluate the visibility, we adopted PSNR as the indicator to reflect the distortion introduced by the embedding operation. The high PSNRs listed in Table 1 indicate that, even if the scheme is applied to the quantized coefficients, the quality degradation is still imperceptible. This can be visually proved by comparing the original image of Barbara in Fig. 6(a) against the watermarked version in Fig. 6(b). We also define the *watermarkable ratio* as the ratio of the number of watermarkable coefficients to the number of all coefficients in X, and the *watermarked ratio* as the ratio of the number of actually modulated coefficients to the number of all coefficients in X. The reason of differentiating these two ratios is because some watermarkable coefficients need not to be modulated if they satisfy Eq. (7). These ratios of each test image are also listed in Table 1. These low ratios demonstrate one of the key features of the proposed scheme: *watermarking only a small proportion of coefficients while providing protection to all of the coefficients and maintaining low distortion.*

5.2 Tamper Detection

A. Cut- and-paste attack

A cut-and-paste attack is to cut a part of the watermarked image and then paste it

somewhere else in the same or another watermarked image. In this experiment, we duplicated the left end of the bookcase in the watermarked image in Fig. 6(b) and attached it to the right end of the bookcase to create a forged image as shown in Fig. 6(c). Fig. 6(d) shows the authentication map D. The solid lines in Fig. 6(d) depict the actual tampered area and are added to illustrate the localization capability of the proposed scheme. Note that because of the DWT decomposition, each coordinate in the $HH1$ band corresponds to a block of 2×2 pixels in the spatial domain. Therefore if one coefficient fails the authentication, the corresponding block of 2×2 pixels in the D map will be turned white to indicate the occurrence of tampering.

B. Vector quantization attack

The idea of vector quantization attack is to forge a new watermarked image (a collage) from a number of authenticated images watermarked with the same secret key by combining portions of different authenticated images while preserving their relative positions in the images. To demonstrate how the proposed scheme provides protection against this attack, we generated four slightly different versions of Mandrill with the difference imperceptible to humans and then watermarked them with the same key. The vector quantization attack is then carried out by taking one quadrant from each of the four watermarked images to form a forged one as shown in Figure 6(a). However, from the authentication map in Figure 6(b), we can clearly see that the image is a forgery. The reason the noises appear only along the borders of the four blocks is that, away from the borders, the contextual dependence with the hierarchical neighbourhood is not disturbed, while along the borders, wrong DWT coefficients of different blocks enters the

hierarchical neighbourhood of the coefficients to be authenticated, which disturb the dependence relationship.

C. Transplantation attack

To demonstrate how the proposed scheme can put up resistance against the transplantation attack, we first created two images as shown in Figure 8(a) and (b). Imagine that the two images have been divided into 32 × 32 blocks of 8 × 8 pixels, with the block at the upper-left corner identified as block (1,1). We have already made the 5 × 5 blocks, starting from block (1, 3), of the image in Figure 8(a) equal to the 5 × 5 blocks, starting from block (6, 3), of the image in Figure 8(b) except that the central block of Figure 8(a) has been stained. We watermarked the two images using the same key, and carried out the transplantation attack on the watermarked version of Figure 8(b) by copying the 3 × 3 (not 5 × 5) blocks, starting from block (2, 4), of the watermarked version of Figure 8(a) to the watermarked version of Figure 8(b), starting from block (7, 4), to generate the attacked image as shown in Figure 8(c). Although the 16 blocks surrounding the 3 × 3 blocks, starting from block (2, 4), in Figure 8(a) and the 3 × 3 blocks, starting from block (7, 4), in Figure 8(b) are exactly the same, they are watermarked differently because different secret sums as defined Eq. (5) were used. As a result, the attack was successfully detected by the proposed scheme as shown in Figure 8(d).

D. Low pass filter

Low-pass filtering is a common operation for removing details from the image. In this

experiment, we applied Gaussian low-pass filtering to the watermarked image of Lena (Fig. 9(a)). The authentication map in Fig. 9 (b) shows that the image has been subjected to global manipulation.

6. CONCLUSIONS

In this work, we proposed a wavelet-based fragile watermarking scheme in attempt to achieve the requirements of high security, low distortion, and high accuracy of tamper localization for authenticating JPEG2000 images. High security is achieved by establishing contextual dependence among coefficients by involving the unwatermarkable coefficients in the creation of watermark and the embedding process. Involving the unwatermarkable coefficients in the creation of watermark and the embedding process is actually an implicit operation of watermarking without physically / explicitly changing those coefficients. This allows the scheme to achieve the requirement of low distortion by watermarking only a small proportion of the DWT coefficients.

7. REFERENCES

[1] S. Chen, H. Leung, "Ergodic chaotic parameter modulation with application to digital image watermarking," *IEEE Trans. on Image Processing*, 14(10), 1590 - 1602 (2005).

[2] C.-T. Li, "Digital fragile watermarking scheme for authentication of JPEG images," *IEE Proceedings - Vision, Image, and Signal Processing*, 151(6), 460 – 466 (2004).

[3] E. Izquierdo, V. Guerra, "An ill-posed operator for secure image authentication,"

IEEE Trans. on Circuits and Systems for Video Technology, 13(8), 842 - 852 (2003).

[4] D. Skraparlis, "Design of an efficient authentication method for modern image and video," *IEEE Trans. Consumer Electronics,* 49(2), 417 – 426 (2003).

[5] Z. M. Lu, D. G. Xu, S. H. Sun, "Multipurpose image watermarking algorithm based on multistage vector quantization," *IEEE Trans. on Image Processing*, 14(6), 822 – 831 (2005).

[6] C. K. Heng and S. Emmanuel, "A finite state transition-based fragile watermarking scheme for JPEG-2000 compressed images," in *Proc. IASTED International Conference on Internet and Multimedia Systems and Applications* (*EuroIMSA*), 615-620 (2005).

[7] H. Ouda, M. R. El-Sakka, "Localization and security enhancement of block-based image authentication," in *Proc. IEEE Int. Conf. Image Process*, I, 673-676 (2005).

[8] J. Tzeng, W.-L. Hwang, I.-L. Chern, "An asymmetric subspace watermarking method for copyright protection," *IEEE Trans. Signal Processing*, 53(2), Part 2, 784 – 792 (2005).

[9] W.-Y. Chen and C.-H Chen, "A robust watermarking scheme using phase shift keying with the combination of amplitude boost and low amplitude block selection," *Pattern Recognition*, 38(4), 587-598 (2005).

[10] K. Su, D. Kundur, D. Hatzinakos, "Spatially localized image-dependent watermarking for statistical invisibility and collusion resistance," *IEEE Trans. on Multimedia*, 7(1), 52 - 66 (2005).

[11] P. S. L. M. Barreto, H. Y. Kim, and V. Rijmen, "Toward secure public-key block-wise fragile authentication watermarking," *IEE Proceedings - Vision, Image*

and Signal Processing, 148(2), 57 – 62 (2002).

[12] J. Fridrich, M. Goljan, and A.C. Baldoza, "New fragile authentication watermark for images," in *Proc. IEEE Int. Conf. Image Processing,* I, 446-449 (2000).

[13] C.-T. Li, D.C. Lou and T.H. Chen, "Image authenticity and integrity verification via content-based watermarks and a public key cryptosystem," in *Proc. IEEE Int. Conf. Image Processing,* III, 694-697 (2000).

[14] M. Yeung and F. Minzter, "Invisible Watermarking for Image Verification," *Journal of Electronic Imaging*, 7(3), 578-591 (1998).

[15] A.J. Menezes, P.C. Van Oorschot, and S.A. Vanstone, Handbook of applied cryptography. CRC Press (1997).

[16] P. W. Wong and N. Memom, "Secret and public key authentication watermarking schemes that resist vector quantization attack," in *Proc. SPIE Security and Watermarking of Multimedia Contents* II*,* 40-47 (2000).

[17] M. Holliman, and N. Memon, "Counterfeiting attacks on oblivious block-wise independent invisible watermarking schemes," *IEEE Trans. Image Processing*, 9(3), 432-411 (2000).

[18] J. Fridrich, M. Goljan, and N. Memon, "Further attack on Yeung-Mintzer watermarking scheme," in *Proc. SPIE Security and Watermarking of Multimedia Contents* II, 428-437 (2000).

[19] C.-T. Li and F.M. Yang, "One-dimensional neighborhood forming strategy for fragile watermarking," *Journal of Electronic Imaging,* 12(2), 284-291 (2003).

[20] M. Wu and B. Liu, "Watermarking for image authentication," in *Proc. IEEE Int. Conf. Image Processing*, II, 437-441 (1998).

[21] L. Xie and G.R. Arce, "A class of authentication digital watermarks for secure multimedia communication," *IEEE Trans. Image Processing,* 10(11), 1754-1764 (2001).

[22] D.A. Winne, H.D. Knowles, D.R. Bull and C.N. Canagarajah, "Digital Watermarking in wavelet domain with predistortion for authenticity verification and localization," in *Proc. SPIE Security and Watermarking of Multimedia Contents* IV, 349-356 (2002).

[23] H. Inoue, A. Miyazaki, and T. Katsure, "Wavelet-based watermarking for tamper proofing of still images," in *Proc. IEEE Int. Conf. Image Processing*, II, 88- 101 (2000).

[24] H. Yuan and X.P. Zhang, "Fragile watermark based on the Gaussian mixture model in the wavelet domain for image authentication," in *Proc. IEEE Int. Conf. Image Processing,* I, 566-569 (2003).

[25] A.H. Paquet, R.K. Ward and I. Pitas, "Wavelet packets-based digital watermarking for image verification and authentication," *Signal Processing,* 83, 2117-2132 (2003).

[26] C. Christopoulos, A. Skodras, and T. Ebrahimi, "The JPEG2000 still image coding system: an overview," *IEEE Transactions on Consumer Electronics*, 46, 1103-1127 (2000).

[27] A.S. Lewis, and G. Knowles, "Image compression using the 2-D wavelet transform," *IEEE Trans. Image Processing,* 1, 244-250 (1992).

Biographies

Chang-Tsun Li received the B.S. degree in electrical engineering from Chung-Cheng Institute of Technology (CCIT), National Defense University, Taiwan, in 1987, the M.S. degree in computer science from U. S. Naval Postgraduate School, U.S.A., in 1992, and the Ph.D. degree in computer science from the University of Warwick, U.K., in 1998. He was an associate professor during 1999-2002 in the Department of Electrical Engineering at CCIT and a visiting professor in the Department of Computer Science at U.S. Naval Postgraduate School in the second half of 2001. He is currently an associate professor in the Department of Computer Science at the University of Warwick, U.K. His research interests include image processing, pattern recognition, computer vision, multimedia security, and content-based image retrieval.

Huayin Si received the B.S. degree in electrical engineering from Beijing Institute of Technology, China in 2003, the MSc degree in computer science from the University of Warwick, UK in 2006. His research interests include image processing, pattern recognition, and computer vision.

Table 1. Embedding distortion measured in PSNR, watermarkable ratio and watermarked ratio.

Image	**Lena**	**Barbara**	**Cameraman**	**Mandrill**
PSNR(dB)	56.52	56.74	55.55	56.62
Watermarkable ratio	21.89%	22.91%	19.22%	24.29%
Watermarked Ratio	15.89%	17.42%	19.51%	21.71%

List of figure captions

Figure1. Incorporation of the proposed watermarking process into the encoder pipeline in JPEG2000 standard. The solid blocks and arrows depict the flow of operations of JPEG2000 standard while the dotted block and arrows indicate the extra watermarking process.

Figure 2. Wavelet-transformed image X, decomposed on n octaves, with $3n + 1$ bands.

Figure 3. Projection operation with value 0 represented in white and 1 in black. One block in C_2 corresponds to 4 child blocks in C_1 while one block in C_3 corresponds to 16 child blocks in C_1. (a) Projecting C_2 to C_1. Note that the two circled '1' blocks in C_2 have no child blocks with value 1 in C_1. Therefore, their values have to be projected down to their child blocks as shown in (b). (b) Projecting C_3 to the *new* C_1. (c) The final C_1

Figure 4. A hierarchical neighborhood centered at the 'black' coefficient, with its members highlighted in gray. Note the black coefficient at the center is also a member of the neighborhood.

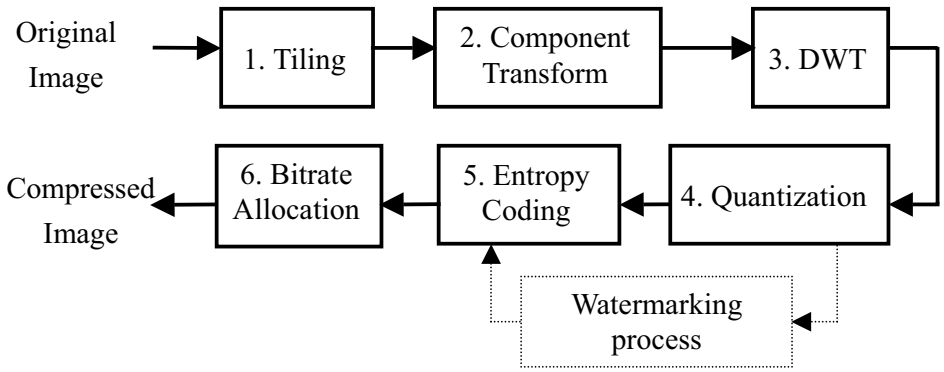
Figure 5: Flowcharts of the proposed scheme. (a) Embedding process (b) Authentication process

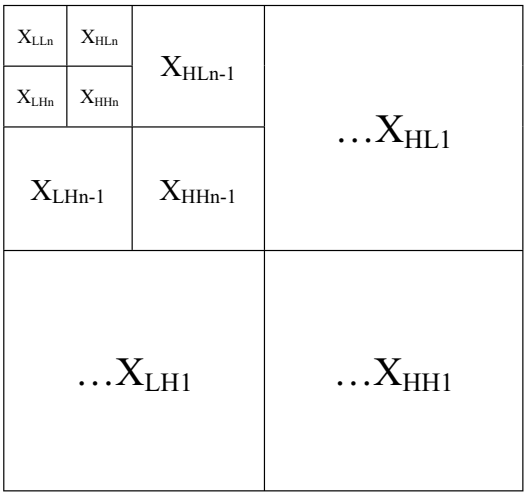
Figure 6. Watermarking and authentication. (a) The original image of Barbara with the quantization step $Q = 5$. (b) The watermarked image of Barbara. The difference between (a) and (b) is imperceptible (PSNR = 56.74 dB). (c) The left end of the bookcase in the watermarked image has been duplicated and attached to the right end. (d) Authentication map indicating the area that has been tampered with. The solid lines depict the actual tampered area and are added to illustrate the localization capability of the proposed scheme.

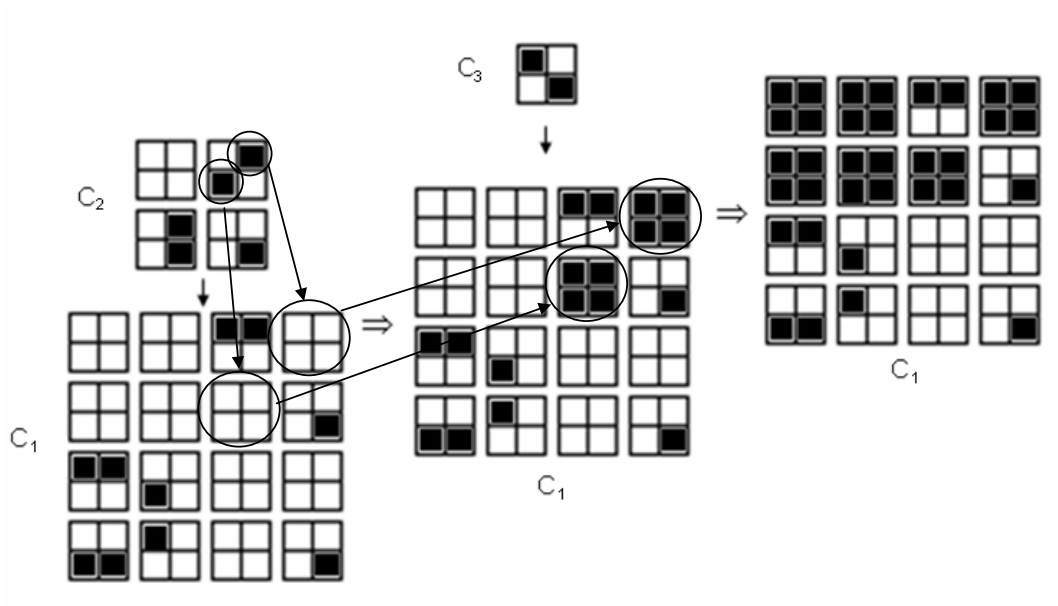
Figure 7. Thwarting vector quantization attack. (a) The forged image with its four quadrants taken from four authenticated images watermarked with the proposed scheme. (b) The authentication map indicating that the image has been subjected to collage / vector quantization attack.

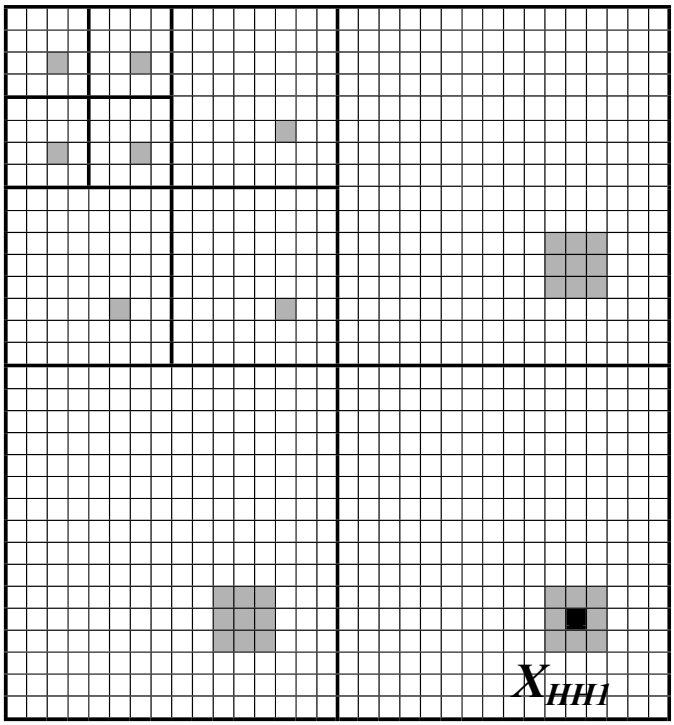
Figure 8. (a) An unwatermarked image. (b) Another unwatermarked image. (c) The watermarked image of Figure 8(b) subjected to the transplantation attack, with the stain on the post copied from a different location in the watermarked version of Figure 8(a). (d) Authentication map indicating the area that has been tampered with.

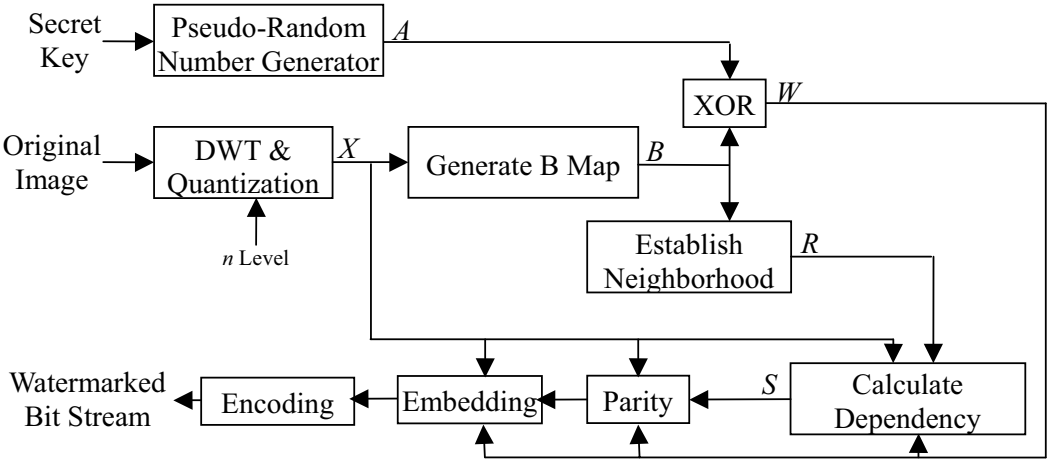
Figure 9. (a) The watermarked image of Lena after low-pass filtering – a form of global attack / manipulation. (b) The authentication map indicating that the image has been subjected to global manipulation.


Original Image
1. Tiling
2. Component Transform
3. DWT
4. Quantization
5. Entropy Coding
6. Bitrate Allocation
Compressed Image
Watermarking process

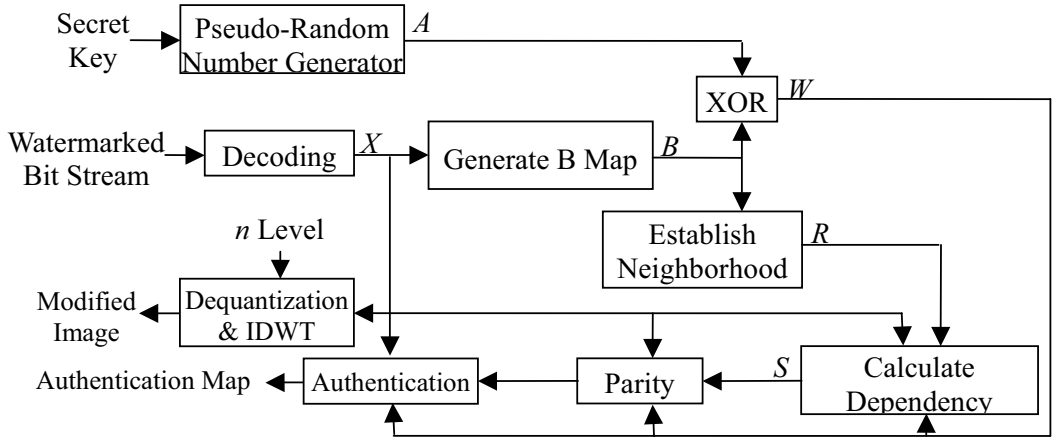

X_{LLn}
X_{HLn}
X_{LHn}
X_{HHn}
X_{HLn-1}
X_{LHn-1}
X_{HHn-1}
…X_{HL1}
…X_{LH1}
…X_{HH1}


C3
C2
C1
C1
C1


X_{HH1}



Embedding process



Authentication process