DUAL-DOMAIN LOCALIZED AUTO-RECOVERY IMAGE AUTHENTICATION

Yassin M. Y. Hasan*, Ammar M. Hassan*
* Electrical Engineering Dept., Assiut University, Assiut, Egypt
E-mail: ymyhasan@aun.edu.eg
** High Institute of Energy, Aswan, Egypt

ABSTRACT - The development of effective image authentication techniques is of remarkably growing interest. Some recently developed fragile, semi-fragile/robust or hybrid-watermarking algorithms not only verify the authenticity of the watermarked image but also provide auto-recovery capabilities. However, several algorithms have been reported as vulnerable to various attacks, especially blind pattern matching (also called copy and vector quantization) attacks with insufficient security. Therefore, compared to existing techniques, we propose a new blind dual-domain self-embedding watermarking scheme with more secure embedding processes of the image’s blocks fragile signatures and robust approximations, more reliable local alterations detection and auto-correction while effectively surviving image content preserving operations.

1 INTRODUCTION

The information systems of the intelligence and police agencies and military require secure extensive storage and manipulation of various types of critical multimedia data. However, the fast progresses of digital technologies have brought many effective tools for copying and modifying the content and ownership information of the various forms of multimedia data. So, recently, not only military and governmental, but also commercial and personal applications recall the use of efficient schemes to protect their interests. Accordingly, multimedia data authentication has become of remarkable interest [1-3]. Direct application of cryptographic authentication schemes may appear as a candidate solution. But, such schemes are not appropriate for multimedia data authentication, because they produce extra authentication bits to be attached to the original data, are not able to localize attacks and are format dependent (for the same audio/image/video content) [4,5]. On the other hand, digital watermarking algorithms are effectively used for this purpose [1-3,6-8].

For image authentication, it is essential that the employed watermarking procedure be blind, secure and so sensitive that minor alterations to the image content be sensed and accurately localized [9,10]. Inserting an undetectable forged mark must be impractical. Many fragile [9-16], semi-fragile [17-19], hybrid fragile-robust [20-28], self-embedding/correction [29-31] watermarking algorithms have lately been introduced for multimedia data authentication.

Fragile watermarking schemes are so sensitive that their watermarks are easily distorted even in case of unsuspicious variations in the image data due to normal image manipulations that do not influence its content [9]. Hence, fragile watermarking is appropriate only in case of lossless environment, i.e., coding, storage, transmission (of the watermarked image). Hence, the primary goal of the attacker facing such fragile watermark is to keep a watermark that makes his tampered or totally fictitious image, “pass” the verification test as authentic [32-34]. This category of attack is basically dissimilar to the attacks against copyright protection and information hiding where the attacker may generally want to considerably damage or cancel the watermark with unnoticeable changes in the image [1-3,6-8]. Several global, block-based and pixel-wise fragile image-watermarking methods have been presented [9-16,35].
On the other hand, semi-fragile watermarking techniques [17-19,28] embed watermarks so robustly to survive (to some, application-dependent, extend) various kinds of typical image processing manipulations such as lossy compression, gamma correction, histogram equalization, linear/nonlinear filtering, scaling, rotation, and mild cropping operations as long as the image content are preserved. At the same time, embedded watermarks must detect malicious alterations.

To gather the different features of fragile-robust/semi-fragile watermarking as needed, few hybrid watermarking algorithms have recently been introduced [29-31]. Furthermore, to not only localize altered regions but also compensate for the damage, self-embedding/auto-correction watermarking techniques have lately been presented that embed an image approximation/visual hash [20-22] into the image itself in a fragile [26,27] or semi-fragile/robust [23-25,28] way using various techniques as given in Section 2.

In this paper, we propose a new blind dual-domain self-embedding watermarking scheme offering more secure embedding processes of the image’s blocks signatures and approximations and more reliable deliberate local alterations detection/auto-correction capabilities, specially compared to [26,27], while surviving image content preserving operations.

2 EXISTING SELF-EMBEDDING WATERMARKING TECHNIQUES

In [24], vectors of similar quantized 8x8 block discrete cosine transform (DCT) coefficients are ordered based on Hilbert curve. Then, a binary watermark is embedded using projections. The paper included a successful example of a region of interest (ROI) self-embedding/reconstruction. In [28], the image is very coarsely approximated using the means of 32x32 blocks. The 8 bits of each (parent) macro-block mean value are embedded in a way that ones modify the middle band DCT coefficients of its 16 8x8 (children) blocks based on a pseudorandom sequence. In [25], a self-authentication/recovery image (SARI) watermarking system is introduced. Noticing that the relationships between the DCT coefficients of block pairs are invariant under JPEG compression, SARI extracts a robust signatures to replace conventional fragile cryptographic hashes. SARI then self-embeds such signatures based on the fact that quantized DCT coefficients are preserved under JPEG quantization with smaller quantization steps. In [26,27], two self-embedding schemes are presented. The 1st one represents each block as 64 bits of 50% JPEG quantized block and inserts them in the LSB of a displaced block. The same displacement vector is employed for all blocks. In the 2nd method, a 16-level (-8 to 7) reduced color-depth image is embedded in the image itself using differential encoding after a fixed cyclic shift of all the image pixels. Four basic weaknesses/security flaws are highlighted in the same presentation [26,27]. Among them, besides its mentioned fragility and attack localization confusion problem, the most critical one is the ease of estimating the locations of all block signatures to forge them.

It is worth to mention that blockwise watermarking schemes offer localized authentication but are vulnerable to blind copy (also called pattern matching and vector quantization) attacks [32-34]. This attack mainly relies on that the watermark embedding and detection processes are run on independent blocks. Once the attacker has a table of authenticated blocks (with the same security parameters), he can use the best-authenticated version approximating an un-authenticated image block without having the verification process detect his introduced falsification. In [35], the authors introduce the use of randomly sized blocks to device a fragile watermarking scheme effectively thwarting such attacks, with improved attack localization.

3 PROPOSED IMAGE AUTHENTICATION ALGORITHM

We propose a novel hybrid block-based watermarking technique that includes both robust self-embedding watermarking scheme for self-correction and fragile watermarking scheme for sensitive authentication. In addition, the proposed hybrid watermarking technique can characterize the detected changes in the watermarked image, specifically, distinguishing between malicious alterations and normal manipulations.

Let the original $N \times M$ image of interest $X$ be divided into $n \times n$ blocks (preferably with $n$ even), with proper padding as needed, to get $M_b$ blocks per row and $N_b$ blocks per column, i.e., $X = \{X_{ij}, i = 0,1,\ldots,N_b-1\}$.
We propose to securely hide an approximated version of \( X' = \{X_{ij}\} \) into the image itself such that each portion of the MSBs of the watermarked image \( X' \) conceals information about other portion(s) of the corresponding approximated image. Furthermore, the LSBs of each block in \( X' \) contains the MSBs’ signatures of the block itself and two other randomly-chosen relatively-distant blocks.

The approximation image can simply be obtained in a blockwise fashion as follows: After excluding the LSBs, each \( n \times n \) block is also equally divided into \( 4 \) sub-blocks and an approximated block version is obtained through retaining only the first \( n_a \) coefficients of the zigzag indexed DCT coefficients of each sub-block. For simplicity, the rest of this paper assumes that \( n=8 \) (compatible with the JPEG standard) and \( n_a=1 \), i.e., only the sub-blocks’ DC coefficients (mean values) be retained. These values are scaled and randomly inserted in \( 4 \) locations of the middle frequency coefficients (shown in Figure 1) of another randomly-selected distant block. Accordingly, the \( 4 \) DC’s values of \( X_{ij} \) are inserted in \( X_{i'j'} \) where indices \( i' \) and \( j' \) are computed as follows:

\[
\begin{align*}
j' &= j + G_{K1}(i) \mod M_b, \quad G_{K1}(i) \in S_{e1} \\
i' &= i + G_{K2}(j) \mod N_b, \quad G_{K2}(j) \in S_{e2}
\end{align*}
\]  

(1)

where \( G_{K1} \) and \( G_{K2} \) denote a keyed pseudo-random number generator (PRNG) with keys \( K1 \) and \( K2 \), respectively, and \( S_{e1} \) and \( S_{e2} \) are sets of pre-selected integer numbers identifying the offsets of the block’s relatively distant region.

Finally, a secure blockwise fragile image authentication algorithm (effectively thwarting blind copy/vector quantization attacks [32-34]) is applied to the self-embedded image. The signature of each block’s MSBs is computed and a doubly-linked chain of inter-block watermark embedding dependency is constructed through embedding, for each block signature \( S_{ij} \), three signature copies \( S_{ij}^0, S_{ij}^1, S_{ij}^2 \) of lengths \( n_p n^2 - 2 \text{int}(n_p n^2/3), \text{int}(n_p n^2/3) \) and \( \text{int}(n_p n^2/3) \) (where \( n_p \) is the number of the LSB planes), respectively, in the LSBs of the block itself and two other randomly-selected relatively distant blocks. Thus, the LSB plane of a block \( X_{ij} \) consists of three signatures in three fields, \( F_{ij}(0), F_{ij}(1) \) and \( F_{ij}(2) \); a signature copy \( S_{ij}^0 \) of the block itself (in \( F_{ij}(0) \)) and signature copies of the block’s randomly-selected two other relatively-distant blocks (in \( F_{ij}(1) \) and \( F_{ij}(2) \)). Specifically, the locations/indices of the two distant blocks relative to \( X_{ij} \) are determined using interleaved pseudo-random row-column cyclic shifts of \( i \) and \( j \) as follows:

\[
\begin{align*}
j_2 &= j_1 + G_{K5}(i_1) \mod M_b, \quad G_{K5}(i_1) \in S_{e5} \\
i_2 &= i_1 + G_{K6}(j_2) \mod N_b, \quad G_{K6}(j_2) \in S_{e6}
\end{align*}
\]  

(2)
where $G_{K3}, G_{K4}, G_{K5}$ and $G_{K6}$ denote a keyed PRNG with keys $K3, K4, K5$ and $K6$, respectively, and $S_{e3}, S_{e4}, S_{e5}$ and $S_{e6}$ are sets of pre-selected integer numbers identifying the offsets of the block’s two relatively distant regions.

Based on the above description and starting with an $N \times M$ image of interest $X$ and a user define key $K = \{ K_i = \text{Rotate}(K,i), i=1,2,\ldots,7 \}$, the overall proposed watermark embedding and detection processes are as follows:

**Embedding Process:** Figure 2 illustrates the proposed self-embedding process that can be summarized in the following steps:

1. **Divide $X$ into $n \times n$ blocks, say $8 \times 8$ blocks.**
2. **For each block $X_{i,j}$:**
   - 1- Sub-divide it into four $(n/2) \times (n/2)$ sub-blocks, say $4 \times 4$ sub-blocks.
   - 2- Compute the DC-coefficients (DCs) of the MSBs of each sub-block.
   - 3- Determine the index $(i_1,j_1)$ of a distant block using (1).
   - 4- Compute the DCT of the MSBs of $X_{i_1,j_1}$ denoted as $X_{i_1,j_1}^{(DCT)}$.
   - 5- Scale the four DCs of the $X_{i,j}$’s sub-blocks by a scale factors $sf$ to yield $X_{i,j}^{(DCs)}$, and make them randomly replace the magnitude of 4 middle frequencies of $X_{i_1,j_1}^{(DCT)}$, but with the same sign, to yield $X_{i_1,j_1}^{(DCT)}$. A keyed PRNG $G_{K7}$ is used for determining the coefficients to be replaced.
   - 6- Compute $X_{i_1,j_1}^{(w)}$ as the rounded inverse DCT (i.e., IDCT) of $X_{i_1,j_1}^{(DCT)}$.
   - 7- If some values of $X_{i_1,j_1}^{(w)}$ are out of range, saturate them to get $X_{i_1,j_1}^{(w)}$.
3. **Assemble $X_{i_1,j_1}^{(w)}$, $ii=0,1,\ldots,N_x-1$, $jj=0,1,\ldots,M_y-1$, to construct the self-embedded image $X^{(w)}$.**

Figure 2. Block diagram of the proposed self-embedding process per block.
d) Apply the proposed blockwise fragile watermarking scheme to \( X^{(w)} \) and concatenate the obtained binary image of block’s signatures as a LSB plane for \( X^{(w)} \) to finally get the authenticated self-embedded image \( X^* \).

**Detection Process:** To verify the authenticity of a given test image \( Z \) (a version of, and hopefully identical to, the watermarked image \( X^* \)), the fragile LSB watermark is tested. Accordingly, the LSBs are excluded and blockwise signatures of the MSBs are computed. Also, using the same user’s key used in the embedding process, the existing 3 versions of each block signature embedded in the LSBs (in the block itself and its 2 randomly-selected distant blocks) are extracted and compared to their corresponding computed signatures. If the computed signature of a block does not match its self embedded signature in LSB’s, blockwise self-reconstruction is employed as summarized in Figure 3 to provide estimates approximating the original block.

![Figure 3. Detection Process: Blockwise self-reconstruction.](image)

Starting with an \( N \times M \) test image \( Z \) and the key \( K \), the overall proposed blockwise watermark extraction/verification and self-reconstruction technique could be summarized in the following steps:

a) Test the fragile watermark to detect and declare the altered blocks.

b) For each altered block \( Z_{i,j} \):
   1. Determine \( i_j \) and \( j_i \), the row/column indices of the block that hides the approximation of \( Z_{i,j} \), using (1).
   2. Extract the block \( Z_{i,j} \).
   3. Compute the DCT of \( Z_{i,j} \) (i.e., \( Z_{i,j}^{(DCT)} \)).
   4. Using \( G_{K^r} \), determine the locations of the four \( DC^* \) coefficients of \( Z_{i,j} \) (i.e., scaled \( DC^* \)s that might have been modified) that are embedded in \( Z_{i,j}^{(DCT)} \) and extract them.
   5. Using the scale factor \( 1/sf \), scale the 4 \( DC^* \)s to get the 4 \( DC^* \)s coefficients of \( Z_{i,j} \).
6-To reconstruct a version of the approximated block, the four DC’s values are rounded, saturated (if overflowed) and interpolated using 2-D bilinear/cubic/spline interpolation to yield an \( n \times n \) (say \( 8 \times 8 \)) block \( \bar{X}_{i,j}^{(R)} \).

c) Assemble the verified blocks with the interpolated versions of the extracted self-embedded approximations of the altered blocks to construct a full estimate \( X^{(R)} \) of the original image.

An overall reconstructed image \( X^{(R)} \) can be obtained using the test image while replacing its (detected) altered blocks with their reconstructed approximations. If there is no block alteration detected, we get \( X = Z = X^{(R)} \), otherwise, given the \( Z \) and \( X^{(R)} \) images, the user can easily identify any local malicious attack. Alternatively, this process can be automated as follows: For each altered block (detected based on the fragile authentication), its four DC values are computed. Then, the correlation, mean square error (MSE) or mean absolute error (MAE) between the re-computed and extracted DC values is evaluated. A local malicious attack usually results in local large deviations between the extracted and re-computed DC values. On the other hand, when the fragile authentication step declares a large number of blocks as altered blocks, the alterations of these blocks may probably be due to normal image manipulations only or both some deliberate attacks combined with normal image manipulations. The spread of the locations and the distribution of the deviations between the extracted and computed DC values, compared to the block activity (edge, texture or uniform), can be used to correctly root the alterations.

It is clear that properly choosing the employed scale factor is very important, because it critically affects the transparency and robustness of the watermark. Hence, two schemes for selecting the employed scale factor(s) are suggested that yield fixed and intra-block adaptive scale factors. The fixed (inter-block adaptive) scale factor is given as follows: \( S_{\text{fixed}} = q_{\text{min}} / D_{C_{\text{max}}} \), where \( q_{\text{min}} \) is the minimum value of the elements of middle frequency region of JPEG standard quantization matrix [36] and \( D_{C_{\text{max}}} \) is the maximum of the DC of the sub-block. To improve the watermark invisibility, an intra-block (per coefficient) adaptive scale factor can be used to scale the \( \overline{D}C \) s of each block by \( q_{l} / D_{C_{\text{max}}} \), where \( q_{l} \) is the value of perceptually optimized DCT (JPEG) quantization matrix [36] for the location chosen using \( G_{k7} \) where the DC watermark value is to be inserted.

4 RESULTS AND DISCUSSION

The proposed method apparently overcomes the weaknesses of the techniques of [26,27] mentioned in Section 2. The performance of the proposed dual-domain image authentication technique has been tested using various types of images. The following examples are considered for illustration using sample images. In these examples, the block size \( n \times n \) and the sets \( S_{\text{cl}}, I=1,2,\ldots,7, \) are chosen to be: \( n = 8 \)

\[
S_{1} = S_{3} = S_{5} = \{ \text{round}(N/3), \text{round}(N/3)+1, \ldots, \text{round}(N/2) \}
\]

\[
S_{2} = S_{4} = S_{6} = \{ \text{round}(M/3), \text{round}(M/3)+1, \ldots, \text{round}(M/2) \}
\]

Example 1: Consider the original \( 512 \times 512 \) 256-garylevel “Girls” image shown in Figure 4 [37]. Applying the proposed hybrid watermark embedding algorithm (using a fixed scale factor \( sf = 1/20 \)), results in the watermarked image of Figure 5, which is perceptually undistinguishable (compared with the original image). The computed peak signal-to-noise ratio PSNR [18] and correlation coefficient between the original and watermarked images are 34.1 dB and 0.9949, respectively.

The watermarked “Girls” image is coded using the JPEG compression standard to test the robustness of the proposed algorithm against it. The performance of the proposed algorithm under the effects of JPEG compression with various compression qualities is illustrated in Figure 6 and Figure 7. Figure 6 shows the correlation between the original watermark (approximated image)/original image and the reconstructed image. Figure 7 shows the PSNR’s of the reconstructed image (for various JPEG quality factors) relative to the original “Girls” image and relative to original watermark at \( sf = 1/10 \), respectively.
Figure 8 shows the PSNR of the reconstructed “Girls” image for various JPEG qualities and various scale factors.

Figure 4. “Girls” image with a marked ROI.

Figure 5. Watermarked image of Figure 4 (with $sf=1/20$) using the proposed hybrid watermarking algorithm. Correlation coefficient=0.9949, PSNR=34.08dB.

Figure 6. Correlation between the reconstructed image (for various JPEG quality factors) and the original image/original watermark at $sf=0.1$.

Figure 7. PSNR of the reconstructed image (for various JPEG quality factors) relative to the original “Girls” image/original watermark at $sf=1/10$. 
Region-of-Interest ROI Authentication with Self Correction: ROI self-correction is proposed to improve the capability of the attack detection/localization and the quality of the reconstructed values for such region(s). For simplicity, we consider a rectangular ROI that can uniquely be specified by the two corners points of its diagonal (cross diagonal). Each bit of the binary representation of the x-y coordinates of these two points are repeated and mapped to +10 or −10 (if 1 or 0, respectively) to construct a vector representing the locations of the ROI. Then, approximated representation is obtained only for the ROI blocks. Both the ROI location information vector and approximated values are robustly embedded only in the non-ROI blocks. Finally, the obtained image is authenticated using the fragile watermark.

Example 2: This example compares the performance of the whole image self-embedding/correction with ROI image self-embedding/correction. Consider the “Girls” host image. The ROI was selected to be the face of one girl, specified by the corner points (288,200) and (504,424). The employed block size is 8×8. The number of ROI (non ROI) blocks is 812 (3248, respectively). When the whole image embedding was used, with a fixed scale factor of 1/20 and sub-block size of 4×4, the watermarked image of Figure 5 was obtained. Under the local attack of Figure 9, the altered blocks were successfully detected with the reconstructed image of Figure 10. When the proposed ROI self-correction was used with a fixed scale factor of 1/5 and sub-block size of 2×2, the watermarked image of Figure 11 was obtained. Under similar local attack, the reconstructed image is shown in Figure 12. Table 1 lists the correlation coefficients between the watermarked and original images, the watermarked ROI and original ROI, and reconstructed ROI and original ROI. It also gives the PSNR’s of the watermarked image/ROI and reconstructed ROI relative to the original image/ROI, respectively. Tables 1 shows that the ROI image self-embedding/correction improves the correlation and PSNR measures of the quality of the watermarked and reconstructed images.
Table 1 Results of the proposed hybrid watermarking with embedding of the whole image approximation versus multiple embedding of the ROI approximation only.

<table>
<thead>
<tr>
<th></th>
<th>Embedding the whole image approximation</th>
<th>Multiple embedding of the ROI approximation only</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Correlation coefficient</td>
<td>PSNR (dB)</td>
</tr>
<tr>
<td>Watermarked image</td>
<td>0.9947</td>
<td>33.960</td>
</tr>
<tr>
<td>Watermarked ROI</td>
<td>0.9923</td>
<td>33.776</td>
</tr>
<tr>
<td>Recon. ROI</td>
<td>0.9780</td>
<td>29.400</td>
</tr>
</tbody>
</table>

5 CONCLUSION

We propose a dual-domain image authentication technique that includes a robust watermarking scheme for self-correction and a fragile watermarking scheme for sensitive authentication. The proposed hybrid authentication technique can characterize the detected changes in the watermarked image to distinguish between malicious manipulations and normal (content preserving) manipulations. Malicious changes of the watermarked image can be corrected. Also, region-of-interest auto-recovery is presented to improve the quality of the watermarked image and the reconstructed values for such region(s) of interest. The proposed method is mainly more secure and effective than existing watermarking schemes such as the methods of [26,27].

REFERENCES