Semi-fragile watermarking based content image authentication scheme

Esquema de autenticación de contenido de imágenes basado en marcas de agua semi-frágil

Clara Cruz, Rogelio Reyes, Mariko Nakano*, Héctor Pérez

Mechanical and Electrical Engineering School, National Polytechnic Institute of Mexico. Av. Santa Ana 1000 Col. San Francisco Culhuacan, CP 04430, México D. F, México

(Recibido el 30 de noviembre de 2009. Aceptado el 18 de mayo de 2010)

Abstract

This paper presents a block-wise semi-fragile watermarking scheme for image content authentication, with tamper area localization and recovery capability. Before the watermark sequence is generated, the original image is divided into two regions: The Region of Interest (ROI) and the Region of Embedding (ROE), where the ROI is selected manually by the owner of the image and the ROE is rest of the whole image after subtracting region belonging to the ROI. The binary watermark of each ROI block is generated by the DC and the 6 lowest AC components in the zig-zag order of the DCT coefficients of these blocks. These extracted bit sequence are embedded into the LSB of the DCT coefficients of the corresponding ROE blocks which are indicated by a mapping list previously generated using a secret key. In the authentication process, the watermark sequence is extracted from the ROI blocks. Then it is compared with the sequence extracted from the corresponding ROE blocks to determine if the ROI block has been altered; in this case the tampered ROI blocks are recovered replacing them by the extracted watermark sequence from ROE. Simulation results show that the reconstructed image quality and the watermark robustness to JPEG compression are better than the previously reported methods by other authors under the same conditions.

----- Keywords: Semi-fragile watermarking, self-correcting, recovery, image authentication, security

Resumen

Este artículo presenta un algoritmo de marca de agua semi-frágil basado en el contenido de los bloques de la imagen para la autenticación de los mismos.

* Autor de correspondencia: teléfono/ fax: 52 + 55 + 565 620 58, correo electrónico: mariko@infinitum.com.mx. (M. Nakano)
Además el algoritmo propuesto tiene la capacidad de recuperar en su forma original los bloques detectados como alterados. Antes de generar la secuencia de marca de agua, la imagen original se divide en dos regiones: la región de interés (ROI) y la región de inserción (ROE), donde la ROI es seleccionada manualmente por el propietario de la imagen mientras que la ROE son todas las regiones de la imagen original fuera de la ROI. La generación de la marca de agua binaria consiste en extraer el coeficiente DC y los 6 primeros coeficientes AC en forma de zig-zag en el dominio DCT de cada bloque ROI. La marca de agua extraída es insertada en el bit menos significativo de los coeficientes DCT de frecuencia media de los 6 bloques ROE, los cuales son seleccionados por medio de una lista de mapeo generada usando una llave secreta. Para autenticar los bloques ROI, se extrae la marca de agua de éstos y de la ROE. Si la diferencia entre ambas secuencias de marca de agua es mayor a un umbral \( Th \), entonces se considera que el bloque fue alterado y se reemplaza con el bloque reconstruido usando la marca extraída de la región ROE. Los resultados muestran una calidad superior de la imagen recuperada así como mayor robustez ante la compresión JPEG del algoritmo propuesto en comparación con los algoritmos propuestos en la literatura, probados bajo las mismas condiciones.

----- Palabras clave: Marca de agua semi-frágil, corrección automática, recuperación, autenticación de imágenes, seguridad

**Introduction**

With the growth of Internet, digital images play an important role to show some evidences in the news and reports in digital media. However, using some software tools, digital images can be easily modified without any traces. Generally these altered images can cause economic and social damages to the involved persons. Therefore the development of a reliable digital image authentication scheme is an urgent issue. Among several approaches, a watermarking based approach is considered as one of alternative solutions. In general, image authentication schemes can be classified into two approaches: digital signature-based authenticators [1, 2] and watermarking-based authenticators [3, 4]. The major difference between these two approaches is that the authentication code is embedded into the same digital media in watermarking-based authenticators, while in the signature-based authenticators it is transmitted or saved separately from digital media. Furthermore the watermarking-based authenticators can be classified into two schemes: fragile watermarking based authentication schemes [3] and semi-fragile watermarking based schemes [4]. The fragile watermarking scheme is used for complete authentication, in which only images without any modification are considered as authentic; while the semi-fragile watermarking schemes can be used for content authentication; in which authenticator distinguish between an image altered intentionally to tamper the image contents, from images suffering content preserving modification, which must be considered as authentic. Therefore the content authentication must be robust to content preserving modification, such as JPEG compression with reasonable compression rate.

Many content authentication methods determine if the image is tampered or not, and some of them can localize the tampered regions [3, 4]; however, very few schemes have the capability to recover the tampered region without using the original image [5-7]. In the authentication and recovery schemes proposed by [5] and [6], the image is divided in 4x4 or 8x8 sub-blocks respectively and...
a mapping list between sub-blocks is generated using the chaotic mixing method with a secret key. In [5], a watermark bits sequence is formed using the output bits of a hash function of the block image data and a cyclic redundancy check bits. While in [6], the watermark bits sequence is formed by a compressed version of an image block, which is extracted by the quantized DCT coefficients. The watermark bits sequence is embedded into two LSB’s of the corresponding image block. From the embedding method and the extracted watermark sequence, both schemes are classified as fragile watermarking scheme, because the embedded watermark sequence is not robust to non-intentional modifications, such as image compression, contamination by noise, etc. In [7], the authors proposed an image authentication scheme with recovery capacity of the modified area, in which the watermark sequence is generated from both the frequency and spatial domains, and embedded into the SPIHT encoded list of significant pixels bit stream. This scheme can be robust to non-intentional modification and its spatial watermark is used in the recovery processing, but the watermark robustness to content preserving modifications is not shown by the authors. In [8], a Slant Transform (SLT) based semi-fragile watermarking scheme for image authentication and self restoration is proposed. Here a pseudorandom watermark sequence and the compressed version of the images are embedded into the image. The watermark sequence is used for authentication and it is embedded into the middle band of SLT, while the compressed version is used for recovery if the image is determined as tampered. The compressed version corresponds to the lower coefficients of SLT, and it is embedded into the LSB in the spatial domain. The principal disadvantage of this scheme is the fragility of the data for recovery due to its embedding domain. In [9] a hybrid block based watermarking technique, which includes robust watermarking scheme for self-correction and fragile watermarking scheme for sensitive authentication, is proposed. In this scheme all alterations, including the content preserving modification, are detected and the recovery mechanism is triggered; therefore the quality of the final recovered image can be affected.

In this paper, an image authentication scheme, with a capability of tampered region localization and recovery, is proposed. In the proposed scheme, an image is segmented in two regions: The Regions of Interest (ROI) and the Regions of Embedding (ROE). The ROI is a region which contains important information that requires some protection, for example regions of faces of persons involved in some scandal scene, while the ROE is rest of the whole image after subtracting region belonged to ROI. Thus the ROE can be, for example, the image background. The information of ROI is encoded to generate a watermark sequence which is embedded into the ROE, of the same image, in an imperceptible manner. In the authentication stage, two watermark sequences, extracted from ROI and ROE respectively, are compared to determine the tampered ROI blocks. If some blocks of ROI are detected as tampered, the recovery process performs a reconstruction of these blocks from the watermark sequence extracted from ROE.

The proposed scheme is evaluated from several points of view: watermark imperceptibility in the watermarked image, accuracy of tampered regions detection, recovery capability of the altered region, quality of the recovered regions and watermark robustness against content preserving modification, such as JPEG compression. Simulation results show a fairly good performance about above four issues. Also the performance of the proposed scheme is compared with previously reported schemes [8, 9], the comparison shows better performance of the proposed scheme than the schemes proposed by [8] and [9].

Proposed authentication method

Watermark sequence generation

Generally in a photo image, some objects or some regions contain more important information than other regions. For example in an image of traffic
accident, perhaps the regions of license plates of vehicles involved in an accident are more important than its background or other vehicles no related with the event. Therefore we define two regions in the image: the region of interest (ROI) and the region of embedding (ROE). The ROI is an important region of the image that requires a protection against malicious modification, while the ROE is the rest of the image that no requires any protection. In the proposed algorithm, information of ROI is extracted to generate a watermark sequence and this sequence is embedded into ROE. Figure 1 illustrates the proposed watermark generation process that can be summarized as follows:

**Figure 1 Watermark sequence generation stage**

1) Subtract 127 from the gray level values of the original image to force the pixel values to be [-127,128]. It reduces DC-coefficient value after the image is transformed by DCT.

2) In the original image $X$, the ROI is selected by owner and automatically the ROE is determined in order that the following condition is satisfied.

$$ROI \cap ROE = \phi \text{ and } ROI \cup ROE = X$$ (1)

3) ROI region is divided into non-overlapping blocks of 8×8 pixels.

4) In each block of ROI, a 66 bits watermark sequence is extracted as follows:

a) Compute the 2D-DCT.

b) The DC-coefficient is rounded and represented by 11 bits (10 bits for absolute value and 1 bit for sign). Because the maximum values of DC for 8x8 blocks of an image with range [-127,128] is 1016, it can be represented in a binary form using 11 bits, including the sign bit.

c) Encode each one of the first 6 lowest frequency AC-coefficients, taking the first 6 AC coefficients in the zig-zag order of the block, with 8 bits together with 1 sign bit (total 9 bits).

5) The watermark sequence length of each ROI block is 66 bits, composed by 11 bits of DC-coefficient, 54 bits corresponding to the 6 AC-coefficients of DCT coefficients and finally we add 1 zero; such that which can be divided into 6 segments with 11 bits sequence per segment.

**Watermark embedding**

The proposed watermark embedding process can be summarized as follows:

1) Using a user’s key $K$, the mapping list between ROI blocks and ROE blocks is constructed.
2) Using this mapping list, each ROI block of 8x8 pixels is mapped into 6 ROE blocks, which are used to embed watermark sequence extracted from the ROI block.

3) In each selected 6 ROE blocks, following processes are carried out.
   a) Apply 2D-DCT to 6 ROE blocks
   b) Quantify them using a quantized matrix $Q$ that corresponds to quality factor 70. This value is selected considering a tradeoff between watermarked image quality and watermark robustness against JPEG compression. Quantization of DCT coefficients by $Q$ is given by:
      \[
      \tilde{C}(u,v) = \left\lfloor \frac{C(u,v)}{Q(u,v)} \right\rfloor
      \]
      where $C(u,v)$ and $\tilde{C}(u,v)$ are the $(u,v)$-th DCT coefficient and it quantized version, respectively, $\left\lfloor x \right\rfloor$ is lower nearest integer value of $x$.
   c) Each 11 bits of watermark sequence is embedded into the LSB of the 11 DCT-coefficients of the middle frequency band of the selected 6 ROE blocks.
   d) The watermarked DCT blocks are multiplied by $Q$.
   e) Finally it is transformed by the inverse DCT to get watermarked blocks.

4) Concatenating all watermarked blocks, the watermarked image is generated.

**Authentication and recovery**

The authentication procedure verifies if the contents of the received image are authentic or not. To authenticate the image, two watermarks must be extracted and then compared. This authentication and recovery process are described as follows:

1) The first watermark $W_{ROI_{ext}}$ is generated from the ROI blocks; these operations are same as the watermark generation process described before.

2) The second watermark $W_{ROE_{ext}}$ is extracted from the ROE blocks. Using the same secret key to construct ROI-ROE mapping lists, the 6 corresponding ROE blocks are determined for each ROI block, from which $W_{ROE_{ext}}$ is extracted.

3) For selected 6 ROE blocks, the following operations are carried out to get $W_{ROE_{ext}}$
   a) Apply 2D-DCT to each one of 6 ROE blocks.
   b) DCT blocks are quantized by quantification matrix $Q$.
   c) 11 bits sequence is extracted from LSB of 11 AC coefficients in the middle frequency band of each ROE block.
   d) Concatenated 6 extracted sequences of longitude 11 bits to generate 66 bits $W_{ROE_{ext}}$.

4) In the watermark comparison between $W_{ROI_{ext}}$ and $W_{ROE_{ext}}$, the tolerant threshold $Th$ is employed to distinguish a content preserving operation from malicious manipulation. This authenticity check is given by (3).

\[
\text{if } \sum \text{XOR}(W_{ROI_{ext}}, W_{ROE_{ext}}) < Th \text{ then the block is authentic}
\]
\[
\text{if } \sum \text{XOR}(W_{ROI_{ext}}, W_{ROE_{ext}}) \geq Th \text{ then the block is modified}
\]

Once the authenticity check indicates that a ROI block was tampered, the recovery process of this ROI block is triggered. The recovery process can be summarized as follows:

1) From the extracted watermark sequence $W_{ROE_{ext}}$, the last bit is eliminated to get a watermark sequence with 65 bits.
2) Assign the first 11 bits of $W_{ROE}^{ext}$ to DC-component and the remaining 54 bits are divided into 6 sequences with 9 bits and these are assigned to 6 lowest AC-coefficients of a recovery DCT block.

3) Compute the inverse 2D-IDCT of the recovery DCT block to get a recovery block.

4) Replace the tampered ROI block by the recovery block.

Experimental results

We conduct three experiments; the first one is to evaluate the watermark imperceptibility. In the second one, the tamper localization accuracy and the recovery capability are evaluated. Finally, in the third experiment, the watermark robustness to incidental modification such as JPEG compression is evaluated. Table 1 lists the values used during these evaluation processes of the proposed scheme.

<table>
<thead>
<tr>
<th>Parameter values used during the evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of test images</td>
</tr>
<tr>
<td>$W_j$</td>
</tr>
<tr>
<td>$Th$</td>
</tr>
</tbody>
</table>

Assuming that the probability density function of an authentic ROI is normal and using 1000 trials we get $N(\mu, \sigma^2) = N(2.2861, 2.2367)$; for a false alarm probability smaller than $10^{-12}$ after some manipulations it follows that $Th = \sqrt{2}erf\left(1 - 2(10^{-12})\right)\sigma + \mu$; thus $Th$ is chosen as 13.

Watermark imperceptibility

Gray-scale images are used for these experiments. Two original images “car” and “camera” are shown by Figure 2(a) and 2(d). In Figures 2(b) and 2(e) the qualities of watermarked images measured by the peak signal to noise ratio (PSNR) between the original image and the watermarked image are 36.8 dB and 33.17 dB, respectively. These results indicate that the image distortions introduced during the watermarking embedding process are not significant. Also, by comparing with the watermarked images shown in Figure 2(b) and 2(e), it follows that it is difficult to distinguish, by the human eye, the difference between the original and watermarked images.

The watermark length of the cover image depends directly on the number of ROI blocks selected by the owner, because per each ROI block (8x8 pixels) we extract 66 bits and need 6 ROE blocks to insert 11 bits per ROE block. Figures 2(c) and 2(f) show an example to the possible ROI blocks selected by the owner of image “car” and “camera” represented by black squares. Figure 3 shows the PSNR values changing number of selected ROI blocks. In this figure we can see that if the number of blocks of interest (ROI) increases, the watermark bit length is also increased and as consequence the quality of the watermarked image is reduced.
Capability of tamper detection and recovery

To evaluate the effectiveness of the proposed authentication scheme, the watermarked images were tampered as shown in Figure 4. In Figure 4(b), the number ‘7’ in the number plate is modified to ‘9’ and in the Figure 4(f), the tower of the background is disappeared. Figure 2(c) and 2(f) show the ROI (black square) and ROE blocks (rest of the image) corresponded to ‘car’ and ‘camera’ images, respectively. The authentication results are presented in figures 4(c) and 4(g), where we can observe that the modified areas were detected and recovered as shown by Figures 4(d) and 4(h), whose PSNR respect to their original version are 43 dB and 39.44 dB, respectively.

Watermark robustness to content preserving modification

Generally any images, including watermarked images, suffer some no-intentional content preserving modifications, such as compression or noise contamination, therefore, watermark robustness against these incidental modifications must be taken into account. In the proposed authentication method, ROI information is embedded as watermark sequence into the quantized DCT coefficients, which is generated by a predefined JPEG quality factor. This embedding method guarantees that a watermark sequence can be extracted in almost intact manner, after watermarked image suffer JPEG compression with a better quality factor than the predefined one. Therefore in the proposed scheme, robustness of the embedded watermark sequence against JPEG compression with a quality factor better than 70 is guaranteed.

Performance comparison

The proposed algorithm is compared with two previously reported algorithms, which are Zhao’s [8] and Hassan’s [9] algorithms, because both algorithms have the recovery capability of the altered region. To realize a fair comparison among these algorithms, the concept of ROI and ROE is adapted to both algorithms. Zhao’s algorithm [8] embeds a binary pseudo-random sequence (20 bits) into the Slant coefficients of ROI blocks, which are used to detect altered ROI blocks. Thus this end, a 64 bits sequence generated from the first 10 Slant coefficients of each ROI block is embedded into the LSB of ROE blocks to recover
the altered regions. Hassan’s scheme [9] embeds a fragile watermark with 20 bits length into the LSB of ROI blocks of the image to detect altered regions. Then four DC coefficients are extracted from four sub-blocks of 4x4 in the DCT domain of ROI blocks, which are used to recover the altered region. These coefficients are embedded into the middle range of the DCT coefficients of the corresponding ROE block.

The performance comparison is carried out from several points of view, such as the recovered image quality, tamper region localization and reconstruction capability, as well as robustness to JPEG compression.

Quality of recovered image

Figures 5(a)-5(f) show a comparison of the recovered images quality using the proposed scheme and the Zhao’s and Hassan’s ones, respectively. Figure 5(a) shows the original image of 512 x 512 pixels and Figure 5(b) shows the ROI blocks represented by black squares and the ROE blocks, which are the rest of the image. Here 169 ROI blocks are selected and the remaining 3927 blocks are determined automatically as ROE blocks. Figures 5(c-e) show the three watermarked images generated by the proposed, the Zhao’s and the Hassan’s schemes together with their respective PSNR values, which are 39.85dB, 41.28 dB and 44.19 dB, respectively. The ROI in the three watermarked images were completely tampered as shown by figure 5(f). Recovered images generated by three algorithms are shown by figures 5(g-i). The PSNRs between the watermarked image and recovered one by the proposed scheme, Zhao’s scheme and Hassan’s scheme are 42.90 dB, 40.37 dB and 39.15 dB, respectively.

Tamper localization and restoration capability

Tamper localization and restoration capability of the tampered region are essential characteristics for this type of watermarking scheme. Figure 6 shows the original ‘car’ image, the image with the selected ROI blocks (indicated by black regions), and the tampered image, in which the number ‘7’ of the license plate is changed by the number ‘9’. Figure 7 shows the comparison of the tampered region localization and restoration capability of three algorithms. Figure 7(a),(c) and (e) show localization accuracy of the tampered regions and the restoration capability of the proposed algorithm, Zhao’s and Hassan’s algorithms, respectively. The PSNR values between the watermarked image and recovered image using these three algorithms are 43.12 dB, 39.41 dB and 37.01 dB, respectively. From these figures, the tampered region detection accuracy of the proposed scheme is better than Zhao’s one, and
it is similar with Hassan’s one. Zhao’s algorithm detects other blocks without any modification as altered blocks. About the restoration capability, the proposed algorithm can recover the altered region with better image quality.

Figure 6 (a) Original image, (b) ROI and ROE blocks and (c) Tampered image

Figure 7 Comparison of localization and restoration capability; (a),(c),(e) show the detected tampered regions and (b),(d),(f) show restoration image of the tampered regions by using the proposed algorithm ((a),(b)), Zhao’s algorithm ((c),(d)) and Hassan’s algorithm ((e),(f))

Robustness to JPEG compression

In order to evaluate the robustness to JPEG compression of the three algorithms, the watermarked images generated by these algorithms were compressed by using the standard JPEG compression varying the quality factors (QF) from 100 to 75. Table 2 shows these comparison results, in which ‘√’ means that the scheme determines authenticity of the image and dB values show the quality of the recovery blocks, which were determinate as tampered.

From the table, the proposed scheme is robust to JPEG compression with QF higher than 80, while Zhao’s scheme is only robust when QF is higher than 95 and Hassan scheme is robust only QF=100. Therefore Zhao’s scheme and Hassan’s one trigger their corresponded restoration process, generating low quality recovered image, when the image is compressed by JPEG with QF is smaller than 95 and 100, respectively.

Conclusions

In this paper, a block-wised image authentication scheme with location and recovery capability is proposed. The watermark embedding and retrieval are performed in a selected DCT frequency band. In the authentication stage, the authenticator compares the watermark sequences extracted from the corresponded ROE blocks with the watermark sequence generated from the ROI blocks, to determine its authenticity. If the difference between these two bits-sequences is higher than the predefined threshold value, the block of a ROI is replaced by the extracted watermark sequence. The proposed algorithm is compared with the Zhao’s and Hassan’s schemes from the recovered image quality, localization and recovery capability as well as the robustness to JPEG compression under the same conditions. Localization accuracy of the proposed algorithm is better than that of Zhao’s algorithm and it is very similar with Hassan’s scheme. In the quality of the recovered image, the proposed algorithm provides better quality compared with other two algorithms. Comparing the watermark size for each ROI block among the three algorithms, the proposed scheme needs only one watermark of 66 bits to authenticate and recover the watermarked image, while other two schemes insert a fragile watermark to authentication process (20 bits; Zhao and Hassan schemes) and 64 bits to recovery process in Zhao’s scheme and 4 integer values in Hassan’s scheme. This means that the proposed algorithm is simpler than other two algorithms in the authentication and recovery process. Also
the proposed method is more robust to JPEG compression (QF=80) than Zhao and Hassan schemes, because the watermark sequence in the proposed method is embedded into the quantified DCT coefficients with quality factor 70, while the watermark sequence for recovery in Zhao’s method and that for authentication in Hassan’s method are embedded using LSB embedding method, whose vulnerability to JPEG compression is well known.

Table 2 Comparison of watermark robustness to JPEG compression

<table>
<thead>
<tr>
<th>QF</th>
<th>100</th>
<th>95</th>
<th>90</th>
<th>85</th>
<th>80</th>
<th>75</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compression rate</td>
<td>4.66 bpp</td>
<td>2.53 bpp</td>
<td>1.95 bpp</td>
<td>1.57 bpp</td>
<td>1.37 bpp</td>
<td>1.20 bpp</td>
</tr>
<tr>
<td>Proposed scheme</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>45.04 dB</td>
</tr>
<tr>
<td>Zhao’s Scheme</td>
<td>√</td>
<td>√</td>
<td>30 dB</td>
<td>28.34 dB</td>
<td>21.47 dB</td>
<td>21.99 dB</td>
</tr>
<tr>
<td>Hassan’s Scheme</td>
<td>√</td>
<td>29 dB</td>
<td>27.8 dB</td>
<td>24.78 dB</td>
<td>20 dB</td>
<td>18.9 dB</td>
</tr>
</tbody>
</table>

References