

# Erased Authentication Watermarking in Binary Document Images

Niladri B. Puhani<sup>1</sup>, Anthony T. S. Ho<sup>2</sup>, F. Sattar<sup>1</sup>

<sup>1</sup> Center for Information Security

School of Electrical and Electronic Engineering  
Nanyang Technological University, Singapore, 639798

<sup>2</sup> Department of Computing

School of Electronics and Physical Sciences  
University of Surrey, Guildford Surrey, UK

E-mail: a.ho@surrey.ac.uk

## Abstract

*In this paper, we propose a new secure authentication method in binary document images using erasable watermarks. For localization, a sufficient number of low-distortion pixels may not be available in a block and to embed the authentication signature with blind detection constraint poses a challenging problem. Also, a perceptual watermark cannot be embedded in white or background regions of the document image, making such regions insecure against hostile attacks. In the proposed method, an erasable watermark is designed and embedded in each block of a document image for localization against tampering and Holliman-Memon attack. After verifying the content of each block, the exact copy of original image can be restored at the blind detector.*

## 1. Introduction

Due to the availability of systems for extensive use of digital data, significant interest in data hiding became perceptible during the last decade. It has become evident that intelligent hiding of a piece of data or information within another digital data could address many practical applications like covert communications, copyright protection and content authentication [1]. There has been a growing interest in the authentication watermarking of binary document images such as text, circuit diagrams, signature, financial and legal documents. For such images, hiding sufficient amount of data for secure authentication without creating visible distortion is difficult. The previous method for localization in binary document images has been reported in [2]. In this method, the original image is

divided into many sub-images and each sub-image is watermarked independently. A two-layer watermark is embedded imperceptibly using a block-wise data hiding technique to verify the integrity of watermarked image and localizing any modification in it. In this method, the size of each sub-image is 128×128 pixels; so its localization accuracy is not satisfactory. The block-wise embedding used in this method also suffers from a parity attack. While imperceptibly embedding the authentication signature in a block, the number of low-distortion pixels should be high and the watermark detection process should be blind. In a reasonable block-size, an insufficient number of low-distortion pixels are available. So we turn our attention to the possibility of embedding the erasable or invertible [3] watermark for localization in binary document images. In the proposed method, a set of suitable pixels with high correlation is found in each block and they are losslessly compressed to construct an erasable watermark.

Authentication schemes using block-wise independent watermarks for localization are vulnerable to the Holliman-Memon attack [4]. In this attack, the attacker creates the forged image by a collage of authentic blocks and the forged image is authenticated. A countermeasure using unique image index has been proposed to resist this particular attack in grayscale images [5]. Recently, a new countermeasure against the Holliman-Memon attack using authenticity measure is proposed for grayscale images [6]. The concept of authenticity measure is of particular interest to this paper and is extended for the case of binary document image authentication. In the next section, we propose a new localization method against tampering and the

Holliman-Memon attack through designing of erasable watermarks. The paper is organized as follows: in Section 2 the new localization method is described. Simulation results and discussions are presented in Section 3. Finally some conclusions are given in Section 4.

## 2. Proposed Localization Method

In this section, we propose a localization method for embedding the authentication signature as an erasable watermark in each block of the document image. In the proposed method we find a set of pixels in each block such that; (1) a high correlation exists among the pixels and the number of such pixels is high (2) the same set of pixels can be found at blind detector and (3) the relevant information is preserved after the embedding process. In a document image, the black pixel whose all neighbor pixels are white is termed as an *isolated* pixel. These pixels are perceived as noise in a sharply-contrasted binary image. A white pixel whose all neighbor pixels are white is termed as a *background* pixel.

The isolated and background pixels do not convey important information within document images. If these pixels are altered, a background noise will be formed in the image. In document images, we obtain information by recognizing various patterns like text, drawing, signature etc. It is known that human vision has remarkable ability to recognize such patterns even in the presence of noise. After embedding an erasable watermark in these pixels, the user can still obtain relevant information about the document. The background pixels occur in long sequences and isolated pixels occur in between them with less probability. Such a set of pixels can be significantly compressed using the run-length coding scheme. Flipping of a background pixel creates an isolated pixel and vice-versa; so blind detection of the embedded pixels is possible.

### Embedding:

1. The original image is divided into non-overlapping blocks of  $40 \times 40$  pixels. Watermarking is performed for each block independently and in a sequential order.
2. In each block, an ordered set of *insignificant* pixels are searched in a sequential scanning order. The following conditions are designed to ensure that after flipping the current pixel, an insignificant pixel should not be detected as a pixel which is not insignificant and vice-versa during blind detection. A

pixel is defined as a pseudo-insignificant pixel, if it satisfies conditions (a) and (b) but does not satisfy condition (c). A pixel is defined to be in the insignificant pixel set if,

- a. The pixel is either a background pixel or an isolated pixel.
- b. In a  $J \times J$  pixel window, there should not be any insignificant or pseudo-insignificant pixel already found in the block.
- c. After flipping the current pixel, there should not be any pixel in its 8-pixel neighborhood which comes before in the scanning order and satisfies the above two conditions.

The minimum value of  $J$  is 5 for achieving correct blind watermark detection. The insignificant pixel set is losslessly compressed using the run-length coding scheme. Let the compressed data be denoted as  $C_D$ .

3. Authentication signature  $S_1$  is computed from the block according to the following equation
$$S_1 = H(C_b, K, I_b, I_K) \quad (1)$$
where  $H$ ,  $C_b$ ,  $K$ ,  $I_b$  and  $I_K$  denote hash function, current block in the original image, secret key, block index and image index, respectively.
4. The message  $m$  which is embedded in the insignificant pixel set producing the watermarked block is constructed. There are three parts concatenated in  $m$ ; (1) the compressed data  $C_D$ , (2) 16-bit image index  $I_K$  and (3) the authentication signature  $S_1$ .
5. The embedding is performed pixel-wise; so an insignificant pixel holds one bit of  $m$  and its pixel value is set equal to the signature bit it holds. Likewise all blocks in the image are watermarked.

### Detection:

1. To verify each block in the test image  $X'$  of size  $M \times N$ , the message  $m'$  is extracted by finding the insignificant pixel set. Its component pieces, the compressed version of insignificant pixel set  $C_D'$ , image index  $I_X^d$  and the signature  $A_S'$  are extracted. The compressed version of the insignificant pixel set together with the current block is used to reconstruct the block  $C_b'$ .
2. The authentication signature  $A_S''$  of the reconstructed block is computed

$$A_S'' = H(C_b', K, I_b, I_X^d). \quad (2)$$

3. A matrix ( $R$ ) of  $(M/40)$  rows and  $(N/40)$  columns is constructed while computing  $A_s^*$  for each block. Each entry of  $R$  represents a particular block in image  $X'$  in a sequential order. The magnitude of an entry in  $R$  is 1 if  $A_s^*$  and  $A_s^*$  exactly match in the corresponding block; otherwise it is equal to 0.
4. Different image indices can be extracted from all the blocks in  $X'$ . Let each such index be termed as the candidate image index ( $I_X^c$ ). For each  $I_X^c$ , the authenticity score ( $A_s$ ) is computed by applying the image index estimation algorithm.
5. The candidate image index with the highest authenticity score is chosen to be the estimated image index, denoted as  $I_E$ . A block in the test image will be declared as authentic if the image index extracted from the block is equal to the estimated image index and the corresponding entry in  $R$  for the block is 1. The authenticity measure ( $A_M$ ) of the test image value is equal to the authenticity score of  $I_E$ .

#### Image Index Estimation Algorithm

For every  $I_X^c$ , a matrix ' $T$ ' of  $(M/40)$  rows and  $(N/40)$  columns is constructed and each entry of  $T$  corresponds to an individual block in a sequential order as found in  $R$ .

if ( $R(u, v) = 1$  and  $I_X^d = I_X^c$ ), then  
      $T(u, v) = 1$   
     else  
      $T(u, v) = 0$   
     end

A score matrix ' $S$ ' of  $(M/40)$  rows and  $(N/40)$  columns is then computed from  $T$ .

if  $T(u, v) = 1$ , then  
      $S(u, v) = (G+1)/B$   
     else  
      $S(u, v) = 0$   
     end

where  $G$  is the total number of 1's in  $T$  that is connected to the present entry at  $(u, v)$  through the 8-connected path and  $B$  is the total number of blocks.

The authenticity score for the candidate image index  $I_X^c$  is,

$$A_s = \frac{1}{B} \sum_{u=1}^{(M/40)} \sum_{v=1}^{(N/40)} S(u, v) \quad (3)$$

### 3. Results and Discussions

In this section, we present simulation results by creating the erasable watermark for our proposed localization method. The authentication signature to be used in this algorithm is the Hashed Message Authentication Code (HMAC). The output 128-bit HMAC is used as the authentication signature and the message ' $m$ ' is constructed for each block. The original and watermarked images are shown in Figure 1 after pixel-wise embedding of the erasable watermark. With no tampering, all blocks in the watermarked images are authenticated. To illustrate the localization capability of the proposed method, the characters in the word 'Primary' at the left and bottom portion of the image are removed and the attacked image is shown in Figure 2. The detection is performed on the attacked image and the inauthentic blocks are shown in the reconstructed image in Figure 2.

We consider two binary document images of size  $480 \times 520$  pixels for demonstrating the effectiveness of the proposed method against the Holliman-Memon attack. The first original image and its corresponding watermarked image are shown in Figure 3. Similarly, the second original image and its corresponding watermarked image are shown in Figure 4. The 16-bit image indices with the decimal equivalent of 48056 and 56273 were used for the first and second images respectively. The fake image was constructed as follows. The equation in the last rows of the second watermarked image was added into the first watermarked image. For this purpose, 8 blocks with the block indices of 146 to 153 in the first watermarked image were replaced with the corresponding blocks in the second watermarked image. The block replacing operation was performed for a total of 8 blocks and the fake image is shown in Figure 5. According to the Holliman-Memon attack, the blocks between the two watermarked images are swapped at identical positions; thus each block of the fake image should be authenticated. The proposed detection method was used to verify each block in the fake image. A total of 8 blocks out of 156 blocks in the fake image were verified to be inauthentic in Figure 5. The estimated image index is 48056 and the authenticity measure of the fake image is approximately 0.9.

The performance of the proposed algorithm can be compared with the previous method [2]. The localization accuracy in the proposed method has been improved to the block-size of  $40 \times 40$  pixels. The possibility of parity attack is not present in the

proposed method because each message bit is embedded in an insignificant pixel instead of a block. The proposed method has the ability to resist the Holliman-Memon attack by using the authenticity measure and image index estimation.

#### 4. Conclusion

In this paper, we proposed a new watermarking method that is useful in localizing content alteration in binary document images using erasable watermarks. The proposed method can localize content alteration in the image with high probability and accuracy. The localization accuracy is significantly improved and the proposed method is capable of withstanding the parity and the Holliman-Memon attack.

#### 5. References

[1] M. D. Swanson, M. Kobayashi and A. H. Tewfik, "Multimedia Data-Embedding and Watermarking Technologies," *Proc. of the IEEE*, vol. 86, no. 6, pp. 1064-1087, 1998.

[2] H. Y. Kim and R. L. de Queiroz, "Alteration-Locating Authentication Watermarking for Binary Images," *Proc. Int. Workshop on Digital Watermarking 2004*, (Seoul), LNCS-2939, 2004.

[3] J. Fridrich, M. Goljan and M. Du, "Invertible Authentication," *Proc. of SPIE, Security and Watermarking of Multimedia Contents*, 2001.

[4] 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.

[5] P.W. Wong and N. Memon, "Secret and Public Key Image Watermarking Schemes for Image Authentication and Ownership Verification," *IEEE Trans. Image Processing*, vol. 10, no. 10, 2001.

[6] Niladri B. Puhan and Anthony T. S. Ho, "Secure authentication watermarking for localization against the Holliman-Memon attack," Accepted in *ACM Multimedia Systems Journal*.

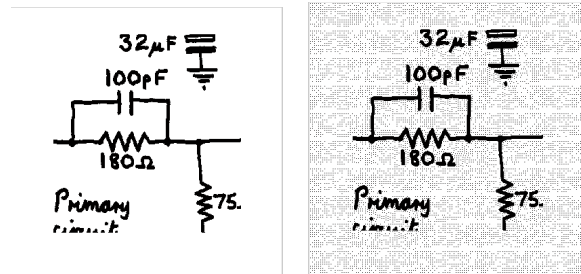


Figure 1: (left) Original image of size 400x400 pixels, (right) watermarked image.

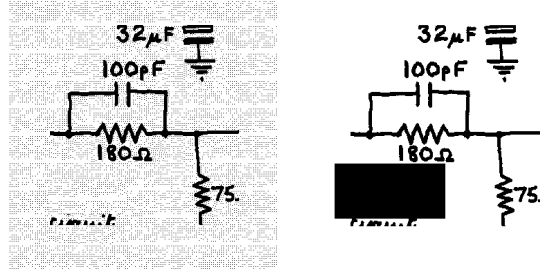


Figure 2: (left) Attacked image, (right) image showing the inauthentic blocks in the dark region.

The authors apply this technique to small 8x8 pixel blocks. The block is DCT transformed, and the frequency masking values  $M(i,j)$  for each frequency bin  $F(i,j)$  are calculated using a frequency masking model. The values  $M(i,j)$  are the maximal changes that do not introduce perceptible distortions. The DCT coefficients are modified to  $F_2(i,j)$  according to the following expression

$$F_2(i,j) = M(i,j) \{ [F(i,j) / M(i,j)] + r(i,j) \text{sign}(F(i,j)) \},$$

where  $r(i,j)$  is a key-dependent noise signal in the interval (0,1), and  $\lfloor x \rfloor$  rounds  $x$  towards zero. Since  $|F(i,j) - F_2(i,j)| \leq M(i,j)$ , the modifications to DCT coefficients are imperceptible.

For a test image block with DCT coefficients  $F_2(i,j)$ , the masking values  $M(i,j)$  are calculated. The error at  $(i,j)$  is estimated by the following equation

The authors apply this technique to small 8x8 pixel blocks. The block is DCT transformed, and the frequency masking values  $M(i,j)$  for each frequency bin  $F(i,j)$  are calculated using a frequency masking model. The values  $M(i,j)$  are the maximal changes that do not introduce perceptible distortions. The DCT coefficients are modified to  $F_2(i,j)$  according to the following expression

$$F_2(i,j) = M(i,j) \{ [F(i,j) / M(i,j)] + r(i,j) \text{sign}(F(i,j)) \},$$

where  $r(i,j)$  is a key-dependent noise signal in the interval (0,1), and  $\lfloor x \rfloor$  rounds  $x$  towards zero. Since  $|F(i,j) - F_2(i,j)| \leq M(i,j)$ , the modifications to DCT coefficients are imperceptible.

For a test image block with DCT coefficients  $F_2(i,j)$ , the masking values  $M(i,j)$  are calculated. The error at  $(i,j)$  is estimated by the following equation

Figure 3: (left) Original image of size 480x520 pixels, (right) watermarked image.

at the quantization level  $l$ . If a wavelet coefficient  $c_l(m,n)$  is chosen for watermark embedding, it is modified so that

$$c_l(m,n) = c_l(m,n) \oplus w(i) \text{ XOR } q(w(i), m,n),$$

where  $w(i)$  is the  $i$ -th watermark bit and  $q(w(i), m,n)$  is a bit generated from the image and a secret key. The construction of the quantization function  $q$  guarantees that one will never have to modify the coefficient at the level  $l$  by more than  $\pm \Delta^l$ . The watermark is extracted by evaluating the expression

$$w(i) = c_l(m,n) \oplus c_l(m,n) \text{ XOR } q(w(i), m,n),$$

at the quantization level  $l$ . If a wavelet coefficient  $c_l(m,n)$  is chosen for watermark embedding, it is modified so that

$$c_l(m,n) = c_l(m,n) \oplus w(i) \text{ XOR } q(w(i), m,n),$$

where  $w(i)$  is the  $i$ -th watermark bit and  $q(w(i), m,n)$  is a bit generated from the image and a secret key. The construction of the quantization function  $q$  guarantees that one will never have to modify the coefficient at the level  $l$  by more than  $\pm \Delta^l$ . The watermark is extracted by evaluating the expression

$$w(i) = c_l(m,n) \oplus c_l(m,n) \text{ XOR } q(w(i), m,n),$$

Figure 4: (left) Original image of size 480x520 pixels, (right) watermarked image.

The authors apply this technique to small 8x8 pixel blocks. The block is DCT transformed, and the frequency masking values  $M(i,j)$  for each frequency bin  $F(i,j)$  are calculated using a frequency masking model. The values  $M(i,j)$  are the maximal changes that do not introduce perceptible distortions. The DCT coefficients are modified to  $F_2(i,j)$  according to the following expression

$$F_2(i,j) = M(i,j) \{ [F(i,j) / M(i,j)] + r(i,j) \text{sign}(F(i,j)) \},$$

where  $r(i,j)$  is a key-dependent noise signal in the interval (0,1), and  $\lfloor x \rfloor$  rounds  $x$  towards zero. Since  $|F(i,j) - F_2(i,j)| \leq M(i,j)$ , the modifications to DCT coefficients are imperceptible.

For a test image block with DCT coefficients  $F_2(i,j)$ , the masking values  $M(i,j)$  are calculated. The error at  $(i,j)$  is estimated by the following equation

$$w(i) = c_l(m,n) \oplus c_l(m,n) \text{ XOR } q(w(i), m,n),$$

The authors apply this technique to small 8x8 pixel blocks. The block is DCT transformed, and the frequency masking values  $M(i,j)$  for each frequency bin  $F(i,j)$  are calculated using a frequency masking model. The values  $M(i,j)$  are the maximal changes that do not introduce perceptible distortions. The DCT coefficients are modified to  $F_2(i,j)$  according to the following expression

$$F_2(i,j) = M(i,j) \{ [F(i,j) / M(i,j)] + r(i,j) \text{sign}(F(i,j)) \},$$

where  $r(i,j)$  is a key-dependent noise signal in the interval (0,1), and  $\lfloor x \rfloor$  rounds  $x$  towards zero. Since  $|F(i,j) - F_2(i,j)| \leq M(i,j)$ , the modifications to DCT coefficients are imperceptible.

For a test image block with DCT coefficients  $F_2(i,j)$ , the masking values  $M(i,j)$  are calculated. The error at  $(i,j)$  is estimated by the following equation

Figure 5: (left) Fake image, (right) detection output; dark region in the image shows the inauthentic blocks.