

# Secure Tamper Localization in Binary Document Image Authentication

Niladri B. Puhon and Anthony T.S. Ho

Center for Information Security, School of Electrical and Electronic Engineering  
Nanyang Technological University, Singapore, 639798  
etsho@ntu.edu.sg

**Abstract.** In this paper, we propose a novel method for secure tamper localization in binary document images using erasable watermarks. For binary images, watermarking with pixel flipping approach is a difficult task, because it can bring noticeable visual distortion. In localization, finding sufficient number of low-distortion pixels in a block to embed the cryptographic signature and their blind detection is more difficult. Also, an imperceptible watermark cannot be embedded in white regions of the document image, making such regions insecure against hostile attacks. In the proposed new method, an erasable watermark is embedded in each block of a document image independently for secure localization. The embedding process introduces some background noise; however the content in the document can be read or understood by the user, because human vision has the inherent capability to recognize various patterns in the presence of noise. After verifying the content of each block, the exact copy of original image can be restored at the blind detector for further analysis. In the proposed method, the tamper localization accuracy is significantly improved as compared to the method proposed by Kim and Queiroz. Simulation results show that an erasable watermark of necessary data length can be embedded in individual blocks of various document images to attain cryptographic security.

## 1 Introduction

Digital watermarking is the art of securing multimedia data by embedding a proprietary mark which may be easily retrieved by the owner of the data to verify about ownership and authenticity. There have been different types of watermarking methods proposed in the literature, designed for various applications [1]. Digital documents and images can be easily and maliciously modified using readily available image processing tools. Thus the need for image authentication methods to establish the integrity and authenticity of digital data is necessary and important. The authentication method is useful in, for example, electronic commerce, legal applications and medical archiving to determine whether the digital data in question has integrity or not. Integrity and authenticity of digital media can be guaranteed through the use of watermarks in conjunction with cryptographic signature. In authentication watermarking, the advantage of having the cryptographic signature hidden inside the digital data rather than appended to it is obvious. Lossless format conversion of the watermarked data does not render it inauthentic though the representation of the data is changed. Another advantage is that if authentication information is localized, then it is possible to achieve the capability to localize the modifications after tampering by a malicious

attacker. Secure fragile watermarking schemes combined with cryptography have been proposed in [2], [3] as a means to verify image authenticity.

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 in which the pixels take on only a limited number of values, hiding significant amount of data for authentication purpose without causing visible artifacts becomes more difficult. As such, many watermarking algorithms use the perceptual model to select low-distortion contour pixels for high watermark capacity [4], [5]. The only method for tamper localization using cryptographic signature in binary document images has been reported by Kim and Queiroz in [6]. 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. The disadvantage of the method is that the size of each sub-image is 128x128 pixels; so its localization accuracy is not good. The block-wise embedding used in this method also suffers from parity attack. The parity attack arises because the signature is embedded by considering the parity of the blocks, i.e. the number of white pixels. If two pixels that belong to the same block change their values, the parity of this block may not change and so this modification will pass undetected. To overcome these shortcomings, we propose an algorithm that is feasible and effective for secure tamper localization in binary documents.

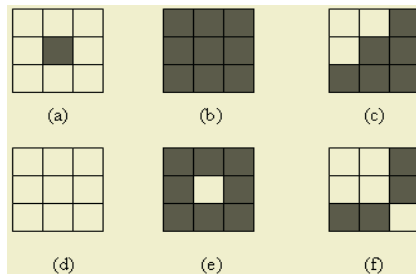
One of the first tamper localization methods was proposed by Wong in [7]. In this scheme, an image is divided into non-overlapping blocks and the watermarking is performed for each block independently. The seven most significant bits of all pixels in a block are hashed using a secure key. The hash output is then XORed with a chosen binary logo and inserted into the LSBs of the same block. At detector, the watermark verification proceeds in the reverse order. However the algorithm suggested by Wong cannot be directly applied in case of binary documents because each pixel has only one bit. By modifying any pixel to embed a watermark would affect the signature of the block and the authentication test would fail. The challenging problem is how to divide a block into two parts such that the above idea of embedding the authentication signature can be applied. While imperceptibly and securely 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 such as 32x32 pixels, there is insufficient number of low-distortion pixels available. The blind detection requirement of these pixels adds to the difficulty in achieving 128-bit watermark capacity for each block. Further, an imperceptible watermark cannot be embedded in white regions of the image. Thus, the white regions are particularly vulnerable to content alteration by the attacker. Due to these shortcomings, it is evident that unless the block size is large, imperceptible watermarking may not be suitable to obtain secure localization in document images.

So we turn our attention to the possibility of embedding the signature in other such pixels that brings visual distortion in watermarked image, but the resulting distortion due to embedding process can be erased entirely at the blind detector. After erasing the embedded watermark, the exact copy of the original image can be restored at the blind detector. This particular concept is known as erasable watermarking in literature and the watermark thus embedded is termed as an erasable watermark. The algorithm

proposed by Fridrich *et al* in [8] for exact authentication of natural images is of particular interest to this paper. Let  $A$  represent the information that is altered in the cover image when we embed a message of  $N$  bits. Fridrich *et al* showed that the erasability is possible provided  $A$  is compressible. If  $A$  can be losslessly compressed to  $M$  bits,  $N-M$  additional bits can be erasably embedded in the cover work. In the implementation of this algorithm for natural images, it is observed that the neighboring pixels are highly correlated. Thus some bit planes in the whole image can be sufficiently compressed to implement an erasable watermark. In case of binary document images each pixel is represented by one bit and it can be considered that there is only one bit plane in the image. If all pixels in the bit plane are losslessly compressed to construct the erasable watermark, then the compressed block does not have perceptual correlation with the original. Therefore, creating an erasable watermark by directly compressing the bit plane is not relevant in document images. Within a block if a set of suitable pixels with high correlation can be found, they can be losslessly compressed and an erasable watermark can be constructed. This motivation is addressed in the next section by proposing a new localization method for binary document images. The paper is organized as follows: in Section 2 the proposed localization method is described. Simulation results and discussion are presented in Section 3. Finally some conclusions are given in Section 4.

## 2 Proposed Localization Method

In this section, we shall propose a localization method for embedding the authentication signature as an erasable watermark in each block of the document image. In the proposed method, after embedding an erasable watermark there will be visible noise in the watermarked image made available for different users. In the embedding process, the relevant information contained in the binary document is preserved so that the user can read or understand the documents. The user can verify the authenticity of such available images and localize any tampering if occurred after watermarking with high probability and accuracy. The watermark can then be erased from the authentic images to retrieve the distortion-free original images for further analysis and application. In the proposed method we find a set of pixels in each block such that; (1) there exists a high correlation 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.



**Fig. 1.** Different categories of pixels in a binary document image based on their 8-neighborhood

As shown in Figure 1 (a)-(f), there can be six categories of pixel neighborhood in such images. Pixels whose neighborhoods have both white and black pixels are contour pixels like in (c) and (f) and they convey maximum information in the document image. The center pixel in (b) is called a *foreground* pixel and the center white pixel in (e) can represent a hole, so these pixels convey some information. The black pixel whose all neighbor pixels are white is termed as an *isolated* pixel like in (a). These pixels are perceived as noise in a sharply-contrasted binary image. As shown in (d), 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 which is similar to the salt-and-pepper noise found in case of natural images. In document images, we obtain information by recognizing various patterns such as symbols, lines and curves etc. It is known that human vision has remarkable ability to recognize such patterns even in the presence of noise. So 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. So 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 in an image; so blind detection of the embedded pixels is possible. We shall outline the proposed method for tamper localization in following steps.

### Embedding

1. The whole image is divided into non-overlapping blocks of  $Y \times Z$  pixels. Watermarking is performed for each block independently and in a sequential order starting from left to right and top to bottom of the image.
2. In each block, an ordered set of *insignificant* pixels are searched in a sequential scanning order. Pixels which are in the border with other blocks are not included in this search to maintain block independence. A pixel is defined to be in the insignificant pixel set if the conditions are satisfied in following order.
  - a) The pixel is either a background pixel or an isolated pixel.
  - b) If there is no other pixel in its 8-neighborhood that has already been decided as an insignificant pixel.
  - c) After flipping, any new insignificant pixel should not be created among already scanned 8-neighborhood pixels.
 Condition (b) is necessary because any pixel previously found to be insignificant shall remain so after its flipping. Condition (c) is necessary because the newly created insignificant pixel will lead to wrong detection.
3. The insignificant pixel set is losslessly compressed using the run-length coding scheme. Authentication signature 'S' is computed from the block according to the following equation.

$$S = H(C_b, K, I_b, I_K) . \quad (1)$$

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.

The block index is used in the computation of signature to resist block-swapping by hostile attacker and the image index is necessary to resist the collage attack [9]. The compressed data and authentication signature are concatenated to create the

message ‘ $m$ ’, which is embedded in the insignificant pixel set. The embedding is performed pixels-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.

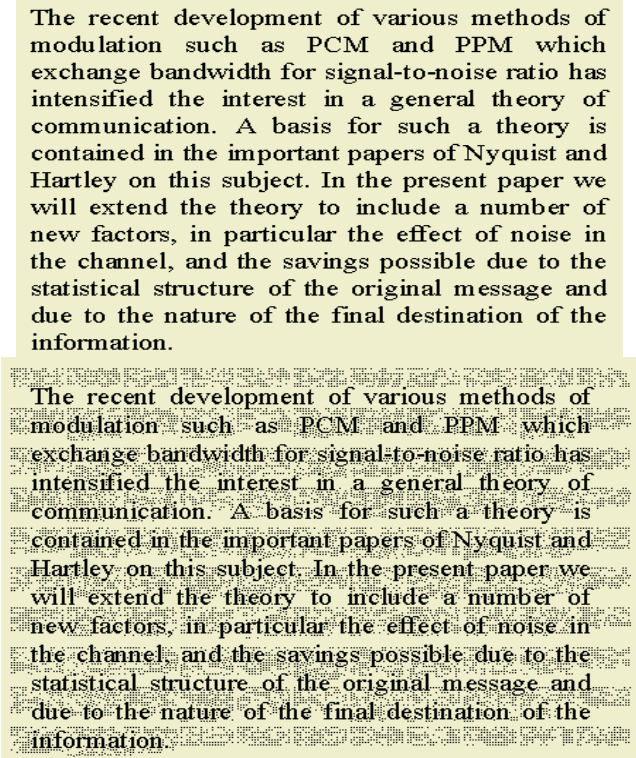
### Detection

4. To verify each block, the message  $m$  is extracted by finding the insignificant pixel set like in steps (1) and (2). Its component pieces, the compressed version of insignificant pixel set and the authentication signature are also extracted. The compressed version of the insignificant pixel set together with the watermarked block is used to reconstruct the original block. The authentication signature of the reconstructed block is computed and compared with the extracted signature.
5. If the two signatures match, the reconstructed block is authentic. Verification of each block is done independently to localize any tampering in the watermarked image.

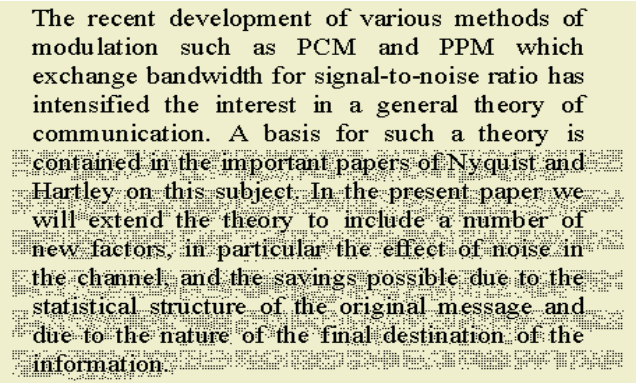
## 3 Results and Discussion

In this section, we present simulation results by creating the erasable watermark for our proposed block-wise localization method. The authentication signature to be used in this algorithm is the Hashed Message Authentication Code (HMAC). The HMAC is found by computing the one way hash function of the data string that is a concatenation of the pixel set and secret key. In our method, provable security against content modification is obtained by using the cryptographic hash function. In the implementation, we have used MD5 [10] hash function to compute the HMAC. The output 128-bit HMAC is used as the authentication signature and the message ‘ $m$ ’ is constructed for each block as described in the proposed method. First ten bits of  $m$  represents the size of the compressed data. While compressing the insignificant pixel set by run-length coding, 10-bit representation is used for the number of white pixels and 1-bit for the number of black pixels. This is because there cannot be two consecutive isolated (black) pixels in the insignificant pixel set. First ten bits and the run-length encoded data represent the compressed data in  $m$ . We have chosen a block size of 32x40 pixels and the block-size can be suitably modified if length of the authentication signature is changed in any case, e.g. 64 or 96 bits.

Figure 2 shows the original and the watermarked image after pixel-wise embedding of  $m$  in each block. It is observed that although background noise is present in the watermarked images, the text can still be read and understood by the user. In the watermarked image, it is observed that the background noise appears to be more random and different well-structured patterns can be recognized due to the inherent ability of human vision. Without any tampering, all blocks in the watermarked images are verified. After verification, the exact copy of original image can be restored at the blind detector. The watermark erasing process is shown in Figure 3. For secure embedding, each block in the original image should have sufficient capacity. The capacity of a block is the number of bits that can be embedded within it. To analyze the performance of the proposed method in different images, we define the term *redundancy* ( $R$ ) in Eq. 2 as number of bits available in a block to accommodate the signature.



**Fig. 2.** Top: original image of 320x440 pixels. Bottom: watermarked image after embedding the erasable watermark in each block

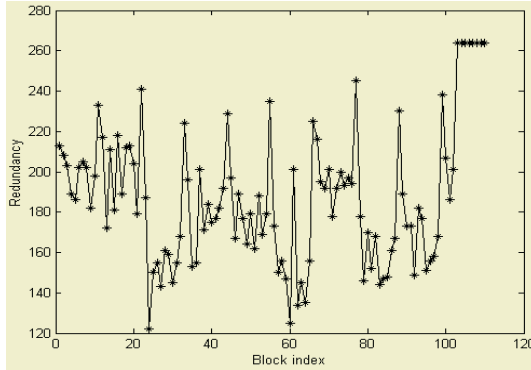


**Fig. 3.** The watermark erasing process is shown in which 44 blocks out of 110 blocks have been restored to its original content

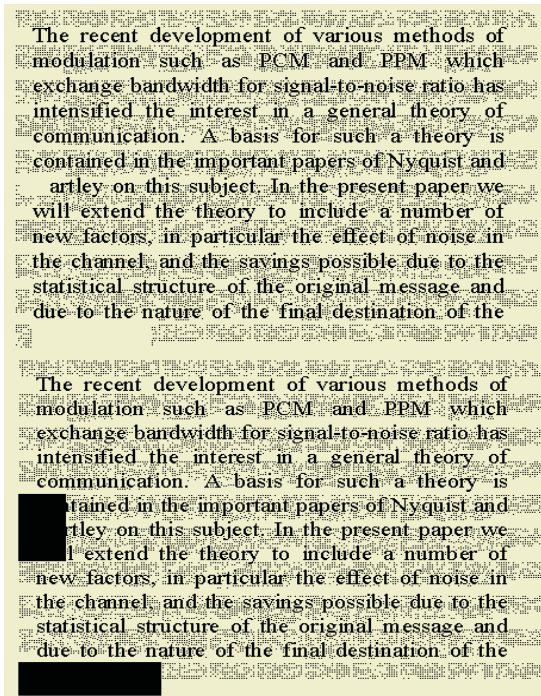
$$R = \text{Size of the insignificant pixel set} - \text{Compressed data size} . \tag{2}$$

In an ideal case, each block of the original image should have  $R \geq 128$  . In Figure 4, it is shown that most blocks in the image have sufficient redundancy for embedding

$m$ . If  $R$  is less than 128 bits in a block, e.g. 120 bits, then first 120 bits of HMAC is used to construct  $m$  and authenticate the current block. Similarly at detector, the comparison between computed and extracted signature is done only for first 120 bits. We perform two alterations in the watermarked image; the first character 'H' in line-7 and the only word 'information.' in last line is deleted as shown in Figure 5. The detection is performed on the attacked image. The detector correctly localizes the tampered blocks in the attacked image.



**Fig. 4.** Redundancy ( $R$ ) for each block of the original image. Except two blocks, all the blocks have  $R \geq 128$



**Fig. 5.** Top: attacked image. Bottom: image showing tamper localization

Though the localization accuracy and cryptographic security offered by the block-wise localization method is high, its block-wise independence was used by Holliman and Memon to design a counterfeiting attack known as collage attack in [9]. To counteract this attack, Wong and Memon suggested including a unique image index while computing the signature in [11]. The use of a unique image index in Eq. 1 for computing the signature removes the possibility of collage attack entirely. Such an approach is feasible for some practical applications; however it may not be always possible because managing such indices brings extra burden to the user. We are currently investigating to suitably extend the proposed method to counteract collage attack in case of binary document images without using image index. To test the effectiveness of proposed method further, a total of 10 binary test images containing text, formulae, drawing and tables are generated. The dimension of the test image is chosen to be the multiple of 32x40 pixel block size. In Table 1,  $R_{mean}$  for each test image is given, where  $R_{mean}$  the mean of redundancy for all the blocks in a test image. The  $R_{mean}$  values for all test images are nearly or above 190 bits. From the Table, it can be shown that using the new method it is possible to securely embed the erasable watermark of necessary length in various document images for tamper localization.

**Table 1.** Redundancy in test images

Number	Size	$R_{mean}$ (bits)
1	480x560	196.62
2	448x520	187.02
3	448x520	190.36
4	480x 480	195.20
5	608x480	188.84
6	480x480	189.80
7	576x480	191.66
8	640x480	205.15
9	608x520	206.91
10	512x520	191.19

The performance of the proposed algorithm can be compared with the previous method [6]. In this method, the localization accuracy was approximately at the block size of 128x128 pixels. In the proposed method it has been significantly improved to approximately the block-size of 32x40 pixels as shown. The block-wise embedding in the previous method suffers from parity attacks as discussed in Section 1. 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 ability of the proposed method for localizing any type of content alteration in the watermarked image is equivalent to the security of cryptographic authentication. In the present method the embedded watermark needs to be erased from the watermarked image for further use, additionally. However, this does not create any bottleneck to the user, because the implementation process is computationally fast. Since the proposed method is relatively simple and no complex perceptual model is necessary, it is possible to obtain fast implementation that is useful for many practical applications of document authentication.

## 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 any kind of content alteration in the image with high probability and accuracy. For this purpose, an ordered set of insignificant pixels was selected and then compressed using the run-length coding scheme. An erasable watermark was constructed by combining the compressed data and HMAC of the block in the original image. After embedding process, the user could easily interpret the document in the presence of noise due to the inherent ability of human vision. After verifying the authenticity, the user could restore the exact copy of the original image for further analysis. The localization accuracy of the proposed method is significantly improved and does not suffer from any parity attack.

## References

1. M. D. Swanson, M. Kobayashi, A. H. Tewfik: Multimedia Data-Embedding and Watermarking Technologies. Proc. of the IEEE, vol. 86, no. 6, (1998)
2. M. Yeung, F. Mintzer: An Invisible Watermarking Technique for Image Verification. Proc. IEEE Int. Conf. Image Processing, Santa Barbara, CA, (1997), 680-683
3. P. W. Wong: A Public Key Watermark for Image Verification and Authentication. Proc. IEEE Int. Conf. Image Processing, Chicago, IL, (1998), 425-429
4. M. Wu, E. Tang, B. Liu: Data hiding in digital binary images. Proc. IEEE Int. Conf. on Multimedia and Expo, Jul 31-Aug 2, (2000), New York
5. A.T.S. Ho, N. B. Puhon, P. Marziliano, A. Makur, Y. L. Guan: Perception Based Binary Image Watermarking. IEEE International Symposium on Circuits and Systems (ISCAS), Vancouver, Canada, 23-26 May (2004)
6. H. Y. Kim, R. L. de Queiroz: Alteration-Locating Authentication Watermarking for Binary Images. Proc. Int. Workshop on Digital Watermarking, Seoul, LNCS-2939, (2004)
7. P. Wong: A Watermark for Image Integrity and Ownership Verification. Proc. IS&T PIC, Portland, Oregon, (1998)
8. J. Fridrich, M. Goljan, M. Du: Invertible Authentication. Proc. of SPIE, Security and Watermarking of Multimedia Contents, (2001)
9. M. Holliman, N. Memon: Counterfeiting attacks on oblivious block-wise independent invisible watermarking schemes. IEEE Trans. Image Processing, vol. 9, no. 3, 432-441, March (2000)
10. R. L. Rivest: RFC 1321: The MD5 Message-Digest Algorithm. Internet Activities Board, (1992)
11. P.W. Wong, N. Memon: Secret and public key image watermarking schemes for image authentication and ownership verification. IEEE Trans. Image Processing, vol. 10, no. 10, October (2001)