# FRAGILE LOGO WATERMARKING FOR PUBLIC AUTHENTICATION

*Sergio Bravo*, Lu Gan*, Asoke K. Nandi* and Maurice F. Aburdene***

*Department of Electrical Eng. & Electronics. The University of Liverpool
Brownlow Hill, Liverpool, L69 3GZ, UK; {sbravo,lugan,aknandi}@liv.ac.uk
**Department of Electrical Engineering. Bucknell University
Lewisburg, PA 17837, USA; aburdene@bucknell.edu

## ABSTRACT

A new fragile logo watermarking scheme is proposed for public authentication and integrity verification of images. The security of the proposed block-wise scheme relies on a public encryption algorithm and a hash function. The encoding and decoding methods can provide public detection capabilities even in the absence of the image indices and the original logos. Furthermore, the detector automatically authenticates input images and extracts possible multiple logos and image indices, which can be used not only to localise tampered regions, but also to identify the original source of images used to generate counterfeit images. Results are reported to illustrate the effectiveness of the proposed method.

***Index Terms***— Image fragile watermarking, authentication, vector quantisation attacks.

## 1. INTRODUCTION

Digital images can be manipulated in such a way that, in some cases, it is difficult, even for trained users, to discern whether an image is genuine. To address this issue, fragile watermarking has been proposed for authentication and integrity verification of digital images. Subsequently, some attacks, such as the vector quantisation (VQ) attacks, have been designed to break the security of some watermarking schemes. Such technique exploits block-wise independence by generating VQ code-books from a set of watermarked images, which are utilised to counterfeit images that would go unnoticed by vulnerable authenticators [1, 2].

A well-known fragile watermarking algorithm was proposed by Wong and Memon in [3], where images are divided into non-overlapping blocks of pixels and then hashed along with a block index and an image index. The resulting bit-streams are XOR-ed with a binary logo, then encrypted with either a secret-key or a public-key algorithm, and finally spread over the least significant bits (LSB) of the pixels in each block. As the block index and image index can establish a block-wise dependence, this algorithm can effectively thwart the VQ attack and overcome the security limitations of previous approaches. However, the detector must be provided with the image index to extract the watermark from a cover image. Furthermore, logos are retrieved for human inspection, but the detector cannot automatically estimate the genuineness of cover images if not provided with the original logos. These restrictions may be impractical for many applications.

Other schemes based on Wong and Memon's work have been proposed in [4, 5, 6]. Celik *et al.* [4] proposed a scheme that involves hierarchically structuring the pixels of the input image. A signature is computed for each level and then embedded in the LSBs of the image in accordance with the hierarchical structure; blocks in lower levels carry a portion of the signatures of upper levels. The scheme is secure against VQ attacks, but the achieved localisation resolution might be deficient for some applications.

Fridrich [5] proposed a block-wise scheme where the authentication data and the information about the origin of the image are separated. Each block is independently hashed, and the resulting bit-stream is XOR-ed with a fixed structure containing an image index, the location of the block, and some extra information (e.g. a serial number). Subsequently, the bit-stream is encrypted and spread over the LSB of the pixels in the block. The scheme is resilient to conventional and VQ attacks, and can be private or public-key depending on the encryption algorithm utilised. Nevertheless, Suthaharan [6] argues that the localisation accuracy achieved by Wong and Memon may be improved without constraining its flexibility, as with the fixed structure-based watermark.

Suthaharan [6] replaced the image index and the encryption algorithm with a master key and a session key used to generate a pattern by performing sequences of geometrical distortions into a gradient image. To retrieve the watermark correctly, the detector must be provided with exactly the same master and session keys as in the embedding process in order to generate an identical pattern. Thus, the session key can be made public, as long as the master key is kept secret. However, each authenticator should be restrained to images watermarked with the same master key, which may be too restrictive for some applications.

In this paper, we propose a new fragile logo watermarking scheme that affords automatic authentication even when the detector is not provided with the image index and the original logo. Besides, the proposed scheme is not only resilient to VQ attacks, but also capable of detecting multiple logos in counterfeit images. The rest of the paper is organised as follows. The proposed scheme and its security analysis are presented in Section 2. In Section 3, some characteristics of the proposed scheme are compared with previous approaches. Finally, some results and conclusions are reported in Sections 4 and 5, respectively.

Sergio Bravo is supported by the National Council of Science and Technology (CONACyT).

## 2. PROPOSED SCHEME

### 2.1. Embedding scheme

An input image $X$, of size $M_X \times N_X$, is divided in non-overlapping blocks of $I_X \times J_X$ pixels. Let $X_q$ denote the $q$-th block, for $q = 1, 2, \ldots, \mathcal{N}_X$, where $\mathcal{N}_X = (M_X N_X)/(I_X J_X)$ is the number of blocks in $X$.

Consider a binary logo $L$ of size $(M_X/I_X) \times (N_X/J_X)$, where $L_q \in \{0, 1\}$ denotes the $q$-th element in the logo. Let $\mathcal{I}_X$ be an image index of $I_X J_X - 1$ bits[1]. Algorithm 1 describes the embedding procedure for each block $X_q$. Note that the system requires a secret key, $K_s$, and an encryption technique which can be either symmetric or asymmetric depending on the application.

**Algorithm 1** Embedding procedure for each block $X_q$.

**Require:** $X_q$, $q$, $K_s$, $\mathcal{I}_X$, $L_q$.
**Ensure:** a watermarked block $X_q^w$.
1: Encode a bit-stream as $S_q = \mathcal{I}_X \,||\, L_q$, where $||$ denotes concatenation.
2: Compute $M_q = \mathcal{H}(\bar{X}_q, q) \oplus S_q$, where $\mathcal{H}(\cdot)$ is a hash function, $\oplus$ denotes the XOR operator, and $\bar{X}_q = \lfloor X_q/2 \rfloor 2$.
3: Set $W_q = Encrypt(M_q, K_s)$.
4: Compute $X_q^w = \bar{X}_q + W_q$.

Finally, all the blocks $X_q^w$ are assembled together to form a watermarked image $X^w$.

### 2.2. Extraction and verification scheme

An input image $Y$, of size $M_Y \times N_Y$, is divided in non-overlapped blocks of $I_Y \times J_Y$ pixels, where $Y_q$ denotes the $q$-th block for $q = 1, \ldots, \mathcal{N}_Y$, and $\mathcal{N}_Y = (M_Y N_Y)/(I_Y J_Y)$. Algorithm 2 describes the steps to retrieve a bit-stream $S'_q$ from a the block $Y_q$ by using a decryption key $K_p$.

**Algorithm 2** Extract the bit-stream $S'_q$ from $Y_q$.

**Require:** $Y_q$, $q$, $K_p$.
**Ensure:** a bit-stream $S'_q$.
1: Set $M'_q = Decrypt(W'_q, K_p)$, where $W'_q = Y_q - \bar{Y}_q$, and $\bar{Y}_q = \lfloor Y_q/2 \rfloor 2$.
2: Compute $H'_q = \mathcal{H}(\bar{Y}_q, q)$.
3: Set $S'_q = H'_q \oplus M'_q$.

Let us commence the verification process by defining $\mathbf{B}$ as the set that contains all the blocks in $Y$, i.e.,

$$\mathbf{B} = \{Y_1, \ldots, Y_{\mathcal{N}_Y}\} \ .$$

Subsequently, split $\mathbf{B}$ into $m$ disjoint subsets $\mathbf{B}_1, \ldots, \mathbf{B}_m$, such that the same bit-stream $\breve{S}_1$ is retrieved from all the blocks in $\mathbf{B}_1$, $\breve{S}_2$ from the blocks in $\mathbf{B}_2$, and so forth.

Consider the case where $Y \simeq X^w$ and $K_p$ is the corresponding public-key of $K_s$; that is, $Y$ is a watermarked, likely altered, image and the correct key is provided to the extraction process. In this scenario, two bit-streams, say $\breve{S}_u$ and $\breve{S}_{u'}$, are expected to be identical except for their LSB. Furthermore, the cardinality of the sets $\mathbf{B}_u$ and $\mathbf{B}_{u'}$ is expected to be much greater than the rest of the subsets, i.e. $\forall_{i \neq u,u'} |\mathbf{B}_u \cup \mathbf{B}_{u'}| >> |\mathbf{B}_i|$. Without loss of generality, let us assume that $\breve{S}_u$ is the bit-stream whose LSB is nought. Thus, if $|\mathbf{B}_u \cup \mathbf{B}_{u'}| > \tau$, where $\tau > 1$ is a predefined threshold, the intensity value of the $q$-th element of a bitmap is encoded as,

$$L'_q = \begin{cases} 0 & \text{if } f(Y_q, K_p) = \breve{S}_u \\ 255 & \text{if } f(Y_q, K_p) = \breve{S}_{u'} \\ 128 & \text{otherwise} \end{cases} \ . \quad (1)$$

Note that tampered blocks are encoded with a different intensity value (128) to ease user inspection. Additionally, the retrieved image index, the $I_Y J_Y - 1$ most significant bits (MSBs) of $\breve{S}_u$, identical to the MSBs of $\breve{S}_{u'}$, is reported to the user.

If the block-size is large enough (see Section 2.3), the cardinality of all the subsets in a non-watermarked, or thoroughly tampered, image is expected to be less than the predefined threshold, i.e. $\forall_i |\mathbf{B}_i| < \tau$. In this case, the detector deems the image as non-watermarked with the key $K_p$ (the authenticator may have been provided with the wrong key).

An especial scenario occurs when more than one pair of bit-streams, say $\breve{S}_u$, $\breve{S}_{u'}$, and, $\breve{S}_v$, $\breve{S}_{v'}$, are pair-wise identical except for their LSB. Additionally, the cardinality of the union of each pair of sets is greater than the threshold, i.e. $|\mathbf{B}_u \cup \mathbf{B}_{u'}| > \tau$ and $|\mathbf{B}_v \cup \mathbf{B}_{v'}| > \tau$. Under these circumstances, a different bitmap can be encoded for each pair of bit-streams by employing Eq. (1). An example of this case is presented in Section 4.

Observe that the proposed method thwarts VQ attacks, as the image index, $\mathcal{I}_X$, and the block index, $q$, prevent the creation of VQ code-books. Furthermore, the system is capable of providing one or more bitmaps for user inspection, as well as a complete report including the integrity of the cover image and the retrieved image indices. All these features can be obtained even without the image index of the received image.

### 2.3. Security analysis

To analyse both the false-negative and the false-positive probabilities, we assume that the output of $\mathcal{H}(\cdot)$ is drawn from a uniform distribution, e.g. in the case of cryptographic hash functions. For the false-negative probability, we define an event $E_1$ as the situation where the bit-stream extracted from a tampered block, $Z_{q_1}$, is identical to the bit-stream originally embedded into the genuine block $X_{q_1}^w$, aside from their LSB, provided that $Z_{q_1} \neq X_{q_1}^w$. The probability that the event $E_1$ occurs is $P_{E_1} = 2^{-(I_Y J_Y - 1)}$. Let $\mathcal{X}_1$ denote the number of occurrences of $E_1$. Thus, we define a false-negative as the event where $T$ blocks are erroneously deemed genuine, even though they have been altered indeed; that is,

$$\mathcal{P}_{fn}(\mathcal{X}_1 = T) = \prod_{i=1}^{T} P_{E_1} = \prod_{i=1}^{T} \frac{1}{2^{I_Y J_Y - 1}} = \frac{1}{2^{T I_Y J_Y - T}} \ . \quad (2)$$

To analyse the false-negative probability, consider a bit-stream, say $\breve{S}_u$, extracted from an arbitrary block, say $Y_{q_1}$, in a non-watermarked image. Let $E_2$ denote the event where

[1]This length has been chosen for sake of simplicity, but the algorithm can easily be modified to support shorter image indices. As a result, less distortion would be induced into cover images at the expense of decreasing the security of the system.

the bit-stream extracted from a different block, say $Y_{q_2}$, is identical to $\breve{S}_u$, save its LSB, provided that $q_1 \neq q_2$. The probability that $E_2$ occurs is $P_{E_2} = 2^{-(I_Y J_Y - 1)}$. Let $\mathcal{X}_2$ denote the number of occurrences of the event $E_2$. Thus, according to the stated in the preceding section, we define a false-positive as the situation where the event $E_2$ occurs more than $\tau$ times in the same non-watermarked image. Hence, we model the false-positive probability as a cumulative binomial distribution [7] given by,

$$\mathcal{P}_{fp}(\mathcal{X}_2 > \tau) = 1 - \sum_{i=0}^{\tau} C_i^{\mathcal{N}_Y} \; P_{E_2}^i (1 - P_{E_2})^{\mathcal{N}_Y - i} \;, \qquad (3)$$

where $C_i^{\mathcal{N}_Y}$ denotes the binomial coefficient, i.e., the possible ordered sequences of $i$ blocks out of $\mathcal{N}_Y$.

## 3. COMPARISON OF CERTAIN CHARACTERISTICS

We briefly describe the general characteristics of the schemes summarised in Table 1, where the proposed scheme exhibits more pervasiveness that the other three.

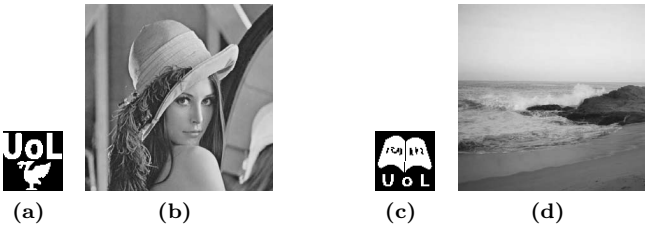
- *Public scheme.* Wong and Memon's [3], Fridrich's [5] and the proposed scheme, can indistinctly use either public-key or private-key encryption methods depending on their applications. Instead of an encryption algorithm, Suthaharan [6] proposed using a master key and a session key to generate a pattern by performing sequences of distortions into a gradient image. However, the same pattern must be generated by the embedding and the extraction methods, and this can only be accomplished by the same master and session keys. As a result, an authenticator should be restrained to images watermarked with the same master key, which may be impractical for public authentication.
- *Logo-based authentication.* Although, strictly speaking, this is not an essential requirement for authentication, *meaningful* watermarks (e.g. logos or seals) can be beneficial for non-technical users [8]. For example, in a judicial enquiry, an extracted logo may help to convince the jury about the original source of a particular image. Except for Fridrich's approach, all the methods shown in the table employ logo-based authentication.
- *Automatic integrity verification.* Most applications would benefit from authentication schemes that automatically detect manipulations in cover images. This is not the case of Wong and Memon's and Suthaharan's schemes, where the integrity verification utterly depends on an error-prone human inspection of the retrieved logo. For these schemes to afford automatic detection, their verification process should be provided with the originally embedded logo, which might be impractical for many applications. On the other hand, neither Fridrich's nor the proposed scheme require the original watermark to verify the integrity of cover images. This characteristic may widen the suitability of the schemes for more applications.
- *Image index-independent extraction.* Suthaharan's approach does not involve image indices, as the block-wise dependency is yield by the pattern generated with a session and a master keys. In Wong and Memon's scheme, the extraction algorithm must be provided with the correct image index of the input image. This restriction significantly reduces the applications of the scheme to those where the user is aware of the correct image index. In contrast, only the correct decryption key is necessary to verify the integrity of a cover image by using either Fridrich's or the proposed approach.
- *Flexible localisation accuracy and distortion boundaries.* The tampering localisation accuracy requirements, as well as the distortion allowed in watermarked images, may vary from application to application. Thus, watermarking schemes with such an adaptability may be suitable for more applications. In Fridrich's scheme, this flexibility is hindered by the very structure of the authentication message, whose parameters cannot be adjusted readily to achieve higher tamper localisation accuracy. On the other hand, the length of the authentication message and the block-size can be simply adapted in the rest of the approaches shown in Table 1, at the expense of increasing/reducing the security of the systems.

**Table 1**. Comparison of the characteristics of the proposed scheme with previous approaches.

| Method | [3] | [5] | [6] | Proposed |
|---|---|---|---|---|
| **Public scheme** | X | X | | X |
| **Logo authentication** | X | | X | X |
| **Auto. verification** | | X | | X |
| **Img. idx.-indep. extraction** | | X | X | X |
| **Flex. localisation and dist.** | X | | X | X |

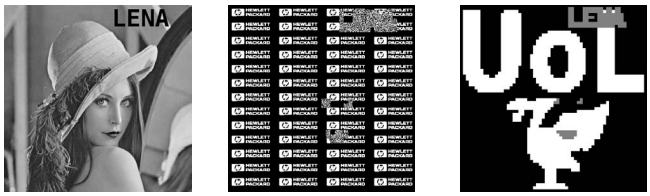
## 4. RESULTS

In our experiments, the logo in Fig. 1(a) was embedded in the standard image of Lena and the output is shown in Fig. 1(b), where the peak signal-to-noise ratio (PSNR) was assessed at 51.1 dB. Similarly, the Beach image, shown in Fig. 1(d) (PSNR $\approx$ 51.2 dB), was watermarked with the logo in Fig. 1(c). We divided the images into blocks of $8 \times 8$ pixels and set the threshold to 40 (approx. 1% of the total blocks). It is also important to mention that the same encryption keys, but different image indices, were utilised to watermark both images.




**(a)** **(b)** **(c)** **(d)**

**Fig. 1**. Watermarked images and embedded logos. (a) and (c) Logos embedded in the images in (b) and (d), respectively.




**(a)** **(b)** **(c)**

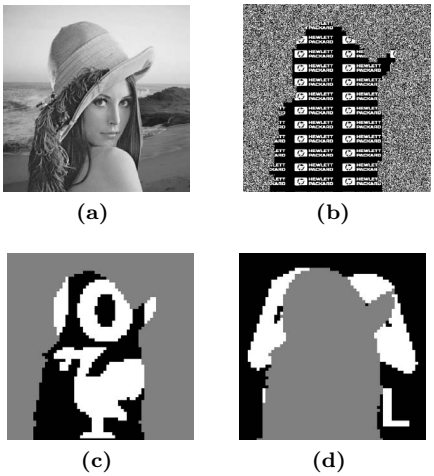
**Fig. 2**. Conventional tampering. (a) tampered image. (b) Logo extracted from image watermarked with Wong and Memon's scheme. (c) Logo extracted from image watermarked with the proposed approach.

To exemplify the tampering detection capabilities of the proposed scheme, the watermarked version of Lena was manipulated with a conventional image editor, as shown in Fig. 2(a). The eyes were slightly lightened, the mouth was darkened, and the legend "LENA" was added in the upper-right corner. Figure 2(c) shows the bitmap encoded by the verification process, which was enlarged to the size of the image to properly localise the tampered regions, which are easily recognisable by the grey colour. Figure 2(b) shows the logo extracted from an identically attacked image watermarked with Wong and Memon's scheme. In this case, the tampered regions can be identified by the noise-like areas.

In the next experiment, we created a counterfeit image by replacing the background of the watermarked version of Lena with a portion of the watermarked version of the Beach image as illustrated in Fig. 3(a). The logos depicted in Figs. 3(c) and 3(d), as well as their respective image index, were recovered by the proposed verification scheme. Thus, the authenticator not only detects the attack, but also provides further information about the source of the images in the counterfeit, e.g., when the same encryption key is used to embed a distinct logo in images captured by cameras that belong to different departments within a company. A similar counterfeit was formed with a pair of images watermarked with Wong and Memon's approach. As in a real scenario, we assumed that the image index of the Lena image is known, whereas the image index of the Beach image is unknown. The bitmap retrieved by this authenticator is shown in Fig. 3(b). Although the manipulation can be easily identified, nothing can be inferred about the other image used in the counterfeit.

## 5. CONCLUSIONS

We have proposed a new fragile logo watermarking scheme for authentication and integrity verification of images. To provide block-wise dependence, and thus resilience to VQ attacks, the method encodes authentication messages dependent on an image index and a block index. Particularly, the verification scheme provides public and automatic detection capabilities even in the absence of the image index and the original logo. Furthermore, possible multiple logos and image indices can be retrieved from counterfeit images to identify the original source of the images used in the attack. Results show that the proposed scheme can detect conventional manipulations as well as sophisticated counterfeits. We are further studying the counterfeiting technique used in the results, as we have discovered that the very purpose of tampering localisation can be easily undermined by similar manipulations in many fragile watermarking methods proposed recently.



**(a)** **(b)** **(c)** **(d)**

**Fig. 3**. Counterfeiting attack. (a) counterfeit image. (b) Logo extracted from image watermarked with Wong and Memon's scheme. (c) and (d) Logos extracted from image watermarked with the proposed scheme.

## 6. REFERENCES

[1] M. Holliman and N. Memon, "Counterfeiting attacks on oblivious block-wise independent invisible watermarking schemes," *IEEE Transactions on Image Processing*, vol. 9, no. 3, pp. 432 – 441, 2000.

[2] J. Fridrich, M. Goljan, and N. Memon, "Further attacks on Yeung-Mintzer fragile watermarking scheme," in *Proceedings of SPIE - The International Society for Optical Engineering*, CA, USA, 2000, vol. 3971, pp. 428 – 437.

[3] P.W. Wong and N. Memon, "Secret and public key image watermarking schemes for image authentication and ownership verification," *IEEE Transactions on Image Processing*, vol. 10, no. 10, pp. 1593 – 1601, 2001.

[4] M. U. Celik, G. Sharma, E. Saber, and A. M. Tekalp, "Hierarchical watermarking for secure image authentication with localization," *IEEE Transactions on Image Processing*, vol. 11, no. 6, pp. 585 – 595, 2002.

[5] J. Fridrich, "Security of fragile authentication watermarks with localization," in *Proceedings of SPIE - The International Society for Optical Engineering*, CA, USA, 2002, vol. 4675, pp. 691 – 700.

[6] S. Suthaharan, "Fragile image watermarking using a gradient image for improved localization and security," *Pattern Recognition Letters*, vol. 25, no. 16, pp. 1893 – 1903, 2004.

[7] Morris H. Degroot and Mark J. Schervish, *Probability and Statistics*, Addison Wesley, 1st edition, 1975.

[8] W. Zeng, B. Liu, and S. Lei, "Extraction of multiresolution watermark images for resolving rightful ownership," in *Proceedings of SPIE - Security and Watermarking of Multimedia Contents*, Apr 1999, vol. 3657, pp. 404–414.