Statistical Fragile Watermarking Capable of Locating Individual Tampered Pixels

Xinpeng Zhang and Shuozhong Wang

***Abstract*—Capability of accurately locating tampered pixels is desirable in image authentication. We propose a novel statistical scheme of fragile watermarking, in which a set of tailor-made authentication data for each pixel together with some additional test data are embedded into the host image. On the authentication side, examining the pixels and their corresponding authentication data will reveal the exact pattern of the content modification. As long as the tampered area is not too extensive, two distinct probability distributions corresponding to tampered and original pixels can be used to exactly identify the tampered pixels.**

***Index Terms*—Fragile watermarking, image authentication, tampered-pixel localization.**

I. Introduction

DIGITAL watermarking techniques that imperceptibly embed additional codes into multimedia for copyright protection have been rapidly developed. While robust watermarks are used for ownership verification, the purpose of fragile watermarks is to check the integrity and authenticity of digital contents [1]. As pirates may try to replace portions of the original content with fake information, it is desirable to be able to locate the modified areas with a fragile watermark.

Many fragile watermarking schemes divide a host image into small blocks and then embed fragile watermark into each block [2], [3]. For example, the embedded watermark may be a hash of the principal content of each 8×8 cover-block. If the image has been tampered, the match between the content and the watermark in the corresponding blocks is destroyed; thus, the tampered blocks can be detected. A smart watermarking method described in [4] uses two pieces of identical index information to generate the fragile watermarks for each block, leading to security against vector quantization attack. This attack selects suitable blocks from many watermarked images to counterfeit an illegitimate image containing a fake complete watermark [5].

Block-wise fragile watermarking schemes have a common limitation. They can only identify the tampered blocks containing modified content but cannot accurately localize the tampered pixels, namely, to find the detailed pattern of the modification. In some applications, however, it is significant to accurately localize the positions of tampered content. For example, if an image is corrupted by salt and pepper noise, accurate localization of the corrupted pixels is useful for identifying the style of the noise and exploiting the preserved content in the image. Fragile watermarking can also be applied in conjunction with image inpainting, which aims to retouch some damaged area by using the surrounding content [6], [7]. Since surviving pixels close to the tampered area provide the most important information for generating the retouched data, it is desirable to accurately localize the tampered area before inpainting.

To this end, some pixel-wise fragile watermarking schemes have been proposed, in which the watermark information derived from the gray values of host pixels is embedded into the host pixels themselves [8], [9]. This way, some tampered pixels can be identified due to the absence of watermark information carried by them. Since some information derived from new pixel values may coincide with the watermark, modification to these pixels cannot be detected directly. In this case, localization of the tampered pixels is not complete, and detection of the tampering pattern is inaccurate.

In this letter, we embed a set of tailor-made authentication data into the host image and incorporate a statistical mechanism to the fragile watermarking scheme. Based on the estimation of the modification strength, two different distributions corresponding to tampered and original pixels can be used to exactly identify the tampered pixels.

II. Statistical Fragile Watermarking Scheme

We first generate a number of authentication bits for each host pixel according to its gray value and embed a folded version of the authentication data and some additional test data into the host image. The folding operation is performed to reduce the necessary host space for accommodating authentication data, and the test data are used to estimate the modification strength. Examination of the pixel values and the authentication data can reveal the trace of any content modification. Although different pixels may have mutual influence in the examination due to the folding, a statistical judging rule will be used to accurately localize the tampered pixels.

A. Watermark Embedding Procedure

Assuming the original image contains N pixels, we denote their gray values as $p_n \in [0, 255]$, $n = 1, 2, \ldots, N$. Each p_n can be represented by 8 bits, $B(p_n, 7), B(p_n, 6), \ldots, B(p_n, 0)$, where

$$B(p_n, u) = \left\lfloor \frac{p_n}{2^u} \right\rfloor \bmod 2, \quad u = 0, 1, \ldots, 7 \tag{1}$$

Manuscript received January 9, 2007; revised February 22, 2007. This work was supported in part by the National Natural Science Foundation of China (No. 60372090 and No. 60502039), in part by the Shanghai Rising-Star Program (No. 06QA14022), and in part by the Key Project of Shanghai Municipality for Basic Research (No. 04JC14037). The associate editor coordinating the review of this manuscript and approving it for publication was Prof. James E. Fowler.

The authors are with the School of Communication and Information Engineering, Shanghai University, Shanghai 200072, China (e-mail: xzhang@shu.edu.cn; shuowang@shu.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LSP.2007.896436

or

$$p_n = \sum_{u=0}^{7} [B(p_n, u) \cdot 2^u]. \tag{2}$$

For each pixel, generate 31 authentication bits as follows:

$$b_{n,t} = \sum_{u=0}^{4} [B(p_n, u+3) \cdot B(t, u)] \bmod 2$$
$$n = 1, 2, \ldots, N; \quad t = 1, 2, \ldots, 31. \tag{3}$$

Here, similar to (1) and (2), we use 5 bits, $B(t,4), B(t,3), \ldots, B(t,0)$, to representing the index t. Equation (3) means that the 31 authentication bits of a pixel are determined by its 5 most significant bits (MSBs). For example, if a gray value is 139, the 18th authentication bit is 1 because $139 = (10001011)_2$, $18 = (10010)_2$, and $1 \times 1 + 0 \times 0 + 0 \times 0 + 0 \times 1 + 1 \times 0 = 1 \bmod 2$, while the 25th authentication bit is 0 because $25 = (11001)_2$ and $1 \times 1 + 0 \times 1 + 0 \times 0 + 0 \times 0 + 1 \times 1 = 0 \bmod 2$.

Property 1: Any alteration on 5 MSBs of a pixel will result in the change of 16 authentication bits.

Proof: Assuming m bits among the 5 MSBs of a pixel p_n are altered ($m = 1$, 2, 3, 4, or 5), the authentication bit $b_{n,t}$ will be changed if and only if the number of $B(t,u)$ with value 1 at the m corresponding positions is odd. Thus, the number of t satisfying this condition is

$$\sum_{\substack{1 \le v \le m \\ v \text{ is odd}}} \binom{m}{v} \cdot 2^{5-m} = 2^{m-1} \cdot 2^{5-m} = 16. \tag{4}$$

So, there are 16 authentication bits changed in total.

The watermark embedding procedure is as follows.

1) For a given host image, calculate all $31 \cdot N$ authentication bits and pseudo-randomly divide them into $(31 \cdot N/11)$ subsets, each of which contains 11 bits, according to a secret key. Then, calculate sums of the 11 authentication bits in each subset with modulus 2, and call the $(31 \cdot N/11)$ folding sums the sum-bits.
2) Further, pseudo-randomly generate $(2 \cdot N/11)$ bits, called the test-bits, according to secret key. The watermark data are made up of the sum-bits and test-bits.
3) Permute the $3 \cdot N$ watermark bits in a pseudo-random way determined by key, and replace the 3 least significant bits (LSBs) of all pixels with them.

Assuming that the original distribution of the 3 LSBs is uniform, the average energy of distortion caused by watermarking on each pixel is

$$E_D = \frac{1}{64} \cdot \sum_{u=0}^{7} \sum_{v=0}^{7} (u - v)^2. \tag{5}$$

So, PSNR is approximately

$$\text{PSNR} \approx 10 \cdot \log_{10} \left(\frac{255^2}{E_D} \right) = 37.9 \,\text{dB}. \tag{6}$$

B. Effect of Tampering

Before introducing the authentication procedure, we first discuss the effect of tampering to the watermarked image. Suppose that an attacker alters the gray values of some pixels without changing the image size. Denote the ratio between the number of pixels with at least one MSB altered and the image size N as r_M, and denote the ratio between the number of LSBs that have been changed and the number of all LSBs $3 \cdot N$ as r_L. Since any alteration on MSB will change 16 authentication bits, there are a total of $16 \cdot r_M \cdot N$ authentication bits being changed. For a subset, its sum-bit will be flipped if the number of changed authentication bits is odd. The probability for this to occur is

$$e = \sum_{\substack{v=1,3,5, \\ 7,9,11}} \left[\binom{11}{v} \cdot \left(\frac{16 \cdot r_M}{31} \right)^v \cdot \left(1 - \frac{16 \cdot r_M}{31} \right)^{11-v} \right]. \tag{7}$$

On the other hand, since the LSBs used for storing the original sum-bits and test-bits are changed with the rate r_L, the probability of a sum-bit being different from the LSB at the corresponding position is

$$E = e \cdot (1 - r_L) + (1 - e) \cdot r_L. \tag{8}$$

Consider a pixel with 5 MSBs unchanged. Its 31 authentication bits, which are also unchanged, are distributed in 31 subsets. In other words, each subset contains one authentication bit and 10 other elements. If the number of changed elements is odd, the sum-bit will be changed. So, the sum-bits of this pixel are changed with probability

$$e_U = \sum_{\substack{v=1,3, \\ 5,7,9}} \left[\binom{10}{v} \cdot \left(\frac{16 \cdot r_M}{31} \right)^v \cdot \left(1 - \frac{16 \cdot r_M}{31} \right)^{10-v} \right]. \tag{9}$$

Then, the probability of the sum-bits being different from the LSBs at the corresponding positions is

$$E_U = e_U \cdot (1 - r_L) + (1 - e_U) \cdot r_L. \tag{10}$$

Denoting the number of sum-bits that do not equal their corresponding LSBs as k_U, it obeys a binomial distribution

$$P(k_U = k) = \binom{31}{k} \cdot E_U^k \cdot (1 - E_U)^{31-k}$$
$$k = 0, 1, \ldots, 31. \tag{11}$$

Furthermore, considering a pixel with at least one MSB altered, there must be 15 authentication bits unchanged and 16 changed. Similarly, the 15 unchanged authentication bits are distributed in 15 subsets. The 15 sum-bits of the subsets will also be changed with the probability given in (9). The probability of the 15 sum-bits being different from LSBs at their corresponding positions is the same as that in (10). We denote the number of sum-bits that do not equal the corresponding LSBs as

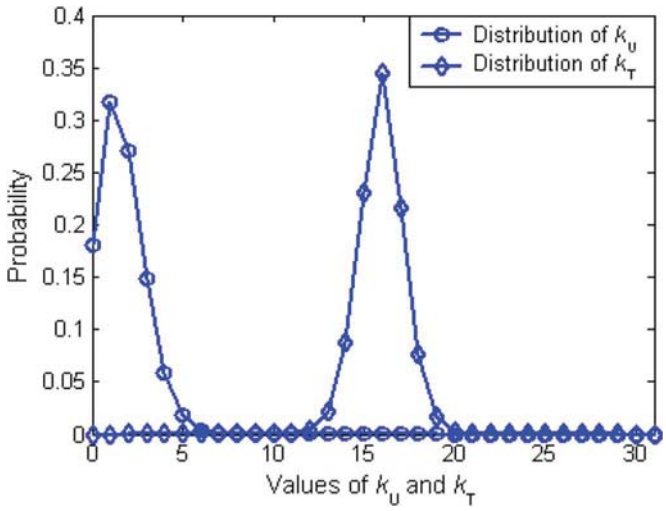

Fig. 1. Distributions of k_U and k_T with $r_M = 0.01$ and $r_L = 0.005$.

k_1. On the other hand, the 16 changed authentication bits correspond to 16 subsets, and their 16 sum-bits will be changed with probability

$$e_T = \sum_{\substack{v=0,2,4,\\ 6,8,10}} \left[\binom{10}{v} \cdot \left(\frac{16 \cdot r_M}{31} \right)^v \cdot \left(1 - \frac{16 \cdot r_M}{31} \right)^{10-v} \right] = 1 - e_U. \tag{12}$$

So, the probability of the 16 sum-bits being different from their corresponding LSBs is

$$E_T = e_T \cdot (1 - r_L) + (1 - e_T) \cdot r_L. \tag{13}$$

We denote the number of sum-bits that do not equal the corresponding LSBs in the 16 sum-bits as k_2. We also denote the number of sum-bits that do not equal the corresponding LSBs in all 31 sum-bits as k_T. Clearly, $k_T = k_1 + k_2$, and its distribution is convolution of the following two binomial distributions:

$$P(k_T = k) = \sum_{v=\max(0,k-16)}^{\min(15,k)} \binom{15}{v} \cdot E_U^v \cdot (1 - E_U)^{15-v} \cdot \binom{16}{k-v} \cdot E_T^{k-v} \cdot (1 - E_T)^{16-k+v} \quad k = 0, 1, \ldots, 31. \tag{14}$$

It can be observed from (11) and (14) that distributions of k_U and k_T are completely different. This of course provides a clue to identify the pixels that have at least one MSB being altered. In fact, when both r_M and r_L are small, the values of e_U and E_U are close to 0, and the values of e_T and E_T are close to 1. According to (11) and (14), the peak of the distribution of k_U is near $k_U = 0$, while that of k_T is at $k_T = 16$. Fig. 1 gives the distributions of k_U and k_T with $r_M = 0.01$ and $r_L = 0.005$.

C. Procedure of Tampered-Pixel Localization

Based on the above discussion, we devise an authentication procedure as follows.

1) For a given image, calculate the $(31 \cdot N/11)$ sum-bits according to its MSBs in the way as given in Step 1 of the watermark embedding procedure, and generate the same $(2 \cdot N/11)$ test-bits according to the secret key.
2) After comparing the calculated sum-bits and the generated test-bits with the LSBs at their corresponding positions, the ratio between the number of different test-bits and $(2 \cdot N/11)$ can be regarded as an estimate of r_L, and the ratio between the number of different sum-bits and $(31 \cdot N/11)$ can be regarded as an estimate of E. According to (7) and (8), an estimate of r_M can be obtained numerically. With the estimates of r_M and r_L, the distributions of k_U and k_T can be found from (11) and (14), respectively.
3) For each pixel, examine its 31 corresponding sum-bits, and count the number of sum-bits different from their corresponding LSBs, k. If

$$(1 - r_M) \cdot P(k_U = k) < r_M \cdot P(k_T = k). \tag{15}$$

this pixel is judged as a tampered pixel, indicating that there is alteration on its 5 MSBs. Note that the original 3 LSBs of all pixels have been replaced with the sum-bits and test-bits; therefore, detection of alteration in the 3 LSB planes is unnecessary.

Consider the two types of false decisions: false negative and false positive. For the former, the number of unaltered original pixels is $(1 - r_M) \cdot N$ and k_U obeys the distribution of (11), whereas the number of tampered pixels is $r_M \cdot N$ and k_T satisfies the distribution of (14). Equation (15) is in fact a MAP criterion that minimizes the total number of false decisions. It indicates that k should be above a threshold T at which the two curves $(1 - r_M) \cdot P(k_U = k)$ and $r_M \cdot P(k_T = k)$ intersect. Let the number of falsely judged pixels be N_F. If r_M and r_L are correctly estimated, the expectation of N_F is

$$E(N_F) = N \cdot (1 - r_M) \cdot \sum_{u=0}^{\lfloor T \rfloor} P(k_T = u) + N \cdot r_M \cdot \sum_{v=\lceil T \rceil}^{31} P(k_T = v). \tag{16}$$

III. Experimental Results

Using a test image Lena sized 512×512 as the host, PSNR due to watermark embedding is 37.9 dB, which verifies the theoretical result in (6) and is imperceptible. We planted a flower on the girl's hat with 2084 pixels replaced. Fig. 2 shows the watermarked image, the tampered version, and the result of tampered-pixel localization. Fig. 2(c) shows a total of 1996 pixels that were correctly judged as "tampered." Because the five MSBs of the rest of the 88 replaced pixels coincided with the original MSBs, they were regarded as "not tampered."

We also replaced pixels in the watermarked image with different rates and then localized the tampered pixels. Denoting the number of replaced pixels N_T, the expectations of r_M and r_L are

$$E(r_M) = \frac{31 \cdot N_T}{32 \cdot 512^2}, \quad E(r_L) = \frac{N_T}{2 \cdot 512^2}. \tag{17}$$

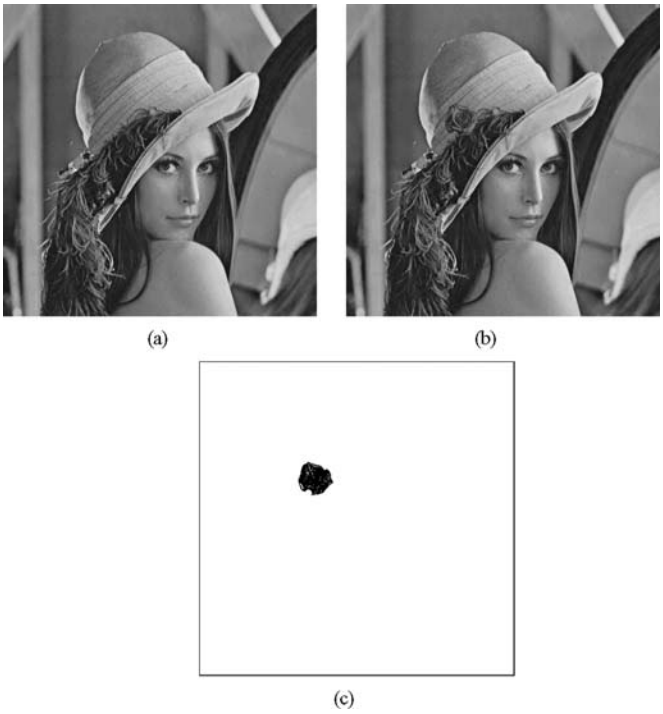

Fig. 2. (a) Watermarked Lena, (b) tampered version, and (c) the tampered-pixel localization result.

Since LSBs of a watermarked image and the substitute are generally more random than the MSBs, r_L was closer to its expectation than r_M in most cases. Fig. 3 gives the experimental results of tampered-pixel localization and the theoretical results calculated from (16). The abscissa represents N_T, and the ordinate is the number of falsely judged pixels N_F. Because there are small differences between the estimated r_M, r_L, and their actual values, the experimental N_F are always slightly larger than the theoretical results. Fig. 3 also shows that judgments for all pixels were correct when the ratio between N_T and the image size was less than 1.1%. Experiments using other host images provided similar results.

IV. Conclusion

This letter proposes a novel fragile watermarking scheme capable of accurately localizing the pixels that are altered in the 5 most significant bits. In this scheme, a folded version of the authentication data derived from MSBs of each pixel and some additional test data used for estimating the tampering strength are

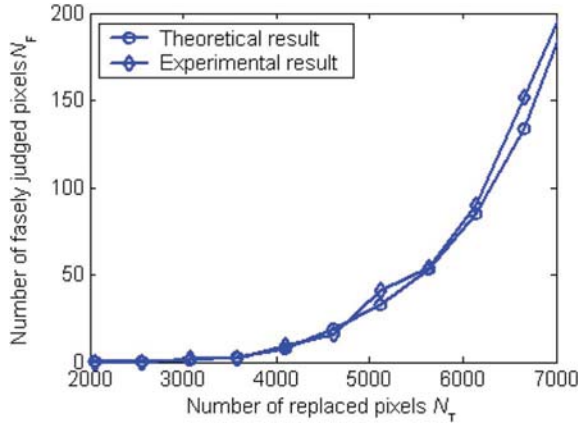

Fig. 3. Comparison between experimental and theoretical results of tampered-pixel localization.

embedded into the host image. If the tampered area is not too extensive, two distinct probability distributions corresponding to the tampered and the original pixels can be used to accurately identify the tampered pixels. Even if the tampered pixels are scattered over the entire image, the statistical mechanism still works, and the tampered pixels can also be located.

References

[1] F. A. P. Petitcolas, R. J. Anderson, and M. G. Kuhn, "Information hiding—A survey," *Proc. IEEE*, vol. 87, no. 7, pp. 1062–1078, Jul. 1999.

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

[3] S. Suthaharan, "Fragile image watermarking using a gradient image for improved localization and security," *Pattern Recognit. Lett.*, vol. 25, pp. 1893–1903, 2004.

[4] J. Fridrich, "Security of fragile authentication watermarks with localization," in *Proc. SPIE, Security and Watermarking of Multimedia Contents IV*, San Jose, CA, Jan. 2002, vol. 4675, pp. 691–700.

[5] M. Holliman and N. Memon, "Counterfeiting attacks on oblivious block-wise independent invisible watermarking schemes," *IEEE Trans. Image Process.*, vol. 9, pp. 432–441, Sep. 2000.

[6] M. Bertalmio, L. Vese, G. Sapiro, and S. Osher, "Simultaneous structure and texture image inpainting," *IEEE Trans. Image Process.*, vol. 12, no. 8, pp. 882–889, Aug. 2003.

[7] A. Criminisi, P. Pérez, and K. Toyama, "Region filling and object removal by exemplar-based image inpainting," *IEEE Trans. Image Process.*, vol. 13, no. 9, pp. 1200–1212, Sep. 2004.

[8] H. He, J. Zhang, and H.-M. Tai, "A wavelet-based fragile watermarking scheme for secure image authentication," in *Proc. 5th Int. Workshop Digital Watermarking*, 2006, vol. 4283, Lecture Notes in Computer Science, pp. 422–432.

[9] S.-H. Liu, H.-X. Yao, W. Gao, and Y.-L. Liu, "An image fragile watermark scheme based on chaotic image pattern and pixel-pairs," *Appl. Math. Comput.*, vol. 185, no. 2, pp. 869–882, Feb. 2007.