Signal Processing 89 (2009) 675–679

Contents lists available at ScienceDirect

Signal Processing

journal homepage: www.elsevier.com/locate/sigpro

Fast communication

Fragile watermarking scheme using a hierarchical mechanism

Xinpeng Zhang *, Shuozhong Wang

School of Communication and Information Engineering, Shanghai University, Shanghai 200072, PR China

ARTICLE INFO

Article history:
Received 14 July 2008
Received in revised form
22 August 2008
Accepted 3 October 2008
Available online 17 October 2008

Keywords:
Fragile watermarking
Tampered-pixel localization
Image restoration

ABSTRACT

This paper proposes a novel fragile watermarking scheme with a hierarchical mechanism, in which pixel-derived and block-derived watermark data are carried by the least significant bits of all pixels. On the receiver side, after identifying the blocks containing tampered content, the watermark data hidden in the rest blocks are exploited to exactly locate the tampered pixels. Moreover, using exhaustive attempts, the proposed scheme is capable of recovering the original watermarked version without any error.

1. Introduction

The purpose of fragile watermarking is to check integrity and authenticity of digital contents and to locate the modified areas by using embedded data. There are two main categories of fragile watermarking techniques: block-wise methods and pixel-wise methods. With the block-wise technique, the host image is divided into small blocks and the mark, e.g., a hash of the principal content of each block, is embedded into the block itself. If an image has been changed, the image content and the watermark corresponding to the tampered blocks are not matched so that the tampered blocks can be detected [1,2]. In general, block-wise fragile watermarking methods are capable of detecting a serious replacement. However, these methods can only identify tampered blocks, but not the tampered pixels. In other words, block-wise fragile watermarking cannot precisely locate the fake content.

Some pixel-wise fragile watermarking schemes have been proposed to solve the problem. The watermark data derived from gray values of host pixels is embedded into the host pixels themselves so that tampered pixels can be identified due to the absence of watermark information they should have carried [3,4]. In these methods, however, since information derived from replaced pixel values may coincide with the watermark itself, localization of the tampered pixels is still inaccurate. In [5], a statistical mechanism is introduced into fragile watermarking, and two different distributions corresponding to tampered and original pixels can be used to exactly locate the tampered pixels. But, this method is effective only when the malicious modification is limited to a small area.

This paper proposes a novel fragile watermarking scheme using a hierarchical mechanism, in which the embedded watermark data are derived both from pixels and blocks. On the receiver side, one can first identify the blocks containing the tampered content, and then use the watermark hidden in the rest blocks to exactly locate the tampered pixels. By combining the advantages of both block-wise and pixel-wise techniques, the proposed scheme is capable of finding the detailed tampered positions even if the modified area is more extensive. Moreover, after localizing the tampered-pixel, the original watermarked version can be perfectly restored using exhaustive attempts.

2. Watermark embedding procedure

In the watermark embedding procedure, the 5 most-significant-bit (MSB) planes in the host image are kept

* Corresponding author. Tel: 86 21 56331551.
E-mail addresses: xzhang@shu.edu.cn (X. Zhang), shuowang@shu.edu.cn (S. Wang).

0165-1684/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.sigpro.2008.10.001

unchanged, and the 3 least-significant-bit (LSB) planes are replaced with watermark data. Here, the watermark data are determined by the MSBs and made up of two parts, which are respectively used to identify tampered blocks and to locate tampered pixels.

The detailed steps are as follows:

1. Denote the numbers of rows and columns in an original image as N_1 and N_2, the total number of pixels as N ($N = N_1 \times N_2$), and the gray pixel-values $p_n \in [0, 255]$, $n = 1, 2, \ldots, N$. Each p_n can be represented with 8 bits, $B_{n,7}$, $B_{n,6}, \ldots, B_{n,0}$, where

$$B_{n,u} = \lfloor p_n / 2^u \rfloor \bmod 2, \quad u = 0, 1, \ldots, 7 \tag{1}$$

For each pixel, generate M authentication bits according to its 5 most significant bits.

$$\begin{bmatrix} a_{n,1} \\ a_{n,2} \\ \\ a_{n,M} \end{bmatrix} = \mathbf{A}_n \cdot \begin{bmatrix} B_{n,7} \\ B_{n,6} \\ B_{n,5} \\ B_{n,4} \\ B_{n,3} \end{bmatrix}, \quad n = 1, 2, \ldots, N \tag{2}$$

where $\mathbf{A}_n$ are pseudo-random binary matrices derived from a secret key, and their size is $M \times 5$. To ensure security, the matrices $\mathbf{A}_n$ should be mutually different. The arithmetic in (2) is modulo-2, meaning that, if there is any change in the 5 MSBs of a pixel, the authentication bits will be flipped with a probability 1/2.

2. According to a secret key, pseudo-randomly divide the $M \cdot N$ authentication bits into a series of subsets, each of which contains K bits. Then, calculate modulus-2 sums of the K authentication bits in each subset, and call the $(M \cdot N/K)$ results the sum-bits. Here, we let M be a multiple of 5 and $K = 2M/5$ so that the number of sum-bits is $5N/2$.

3. Assuming that both N_1 and N_2 are multiples of 8, we divide the original image into $N/64$ non-overlapped blocks sized 8×8. In each block, we pseudo-randomly select 160 positions from the 3 LSB-layers according to the secret key. Also, the LSB-selection in different blocks should be mutually different. Then, a total number of selected LSB is $5N/2$, and replace the original bits at the selected positions with the sum-bits.

4. For each block, we collect the 320 original bits in the 5 MSB-layers and the 160 sum-bits used to replace the selected LSBs. Then, feed the 480 bits into a hash function to compute 32 hash-bits. Here, the hash function must have the property that any change on an input would result in a completely different output. Put the hash-bits into the 32 remaining positions in the 3 LSB-layers, and combine the original MSBs and the substituted LSBs to produce a watermarked image.

The procedure of watermark embedding is sketched in Fig. 1.

3. Tampered-pixel localization and restoration

Assume that an attacker may alter the gray values of some pixels without changing the image size. After receiving the image, we want to locate the tampered pixels and restore the original content. Here, "tampered pixels" are those with changes in their 5 MSBs.

The tampered-pixel localization procedure is made up of two stages. The first is to identify the tampered blocks. After dividing the received image into non-overlapped 8×8 blocks, we select 160 positions from the 3 LSB-layers in each block according to the same secret key. For each block, if the hash result of the 320 bits in the 5 MSB-layers and the 160 bits at the selected positions in 3 LSB-layers is identical to the 32 bits at the other LSB positions, the block is judged as "not tampered". Otherwise, a "tampered" decision is made, meaning that some content in the block has been modified. Here, a block without any modification must be judged as "not tampered", and the probability with which a block containing modified contents is falsely judged as "not tampered" is 2^{-32}. Cases of false judgments are therefore negligible because of the extremely low probability.

In the second stage, we locate the tampered pixels in the tampered blocks. Denote a ratio between the numbers of tampered blocks and that of all blocks as α. Considering a pixel in tampered blocks, its M authentication bits are distributed into M subsets, each of which contains K elements. For each subset, if all the other $(K-1)$ authentication bits in it are derived from pixels in "not tampered" blocks and its sum-bit is also hidden in a "not tampered" block, we say the subset is usable for the pixel. So, a receiver can derive the $(K-1)$ authentication bits in usable subset from their corresponding pixels, and extract the sum-bit from its embedding position. The probability of a subset being usable for a certain pixel is

$$p_U = (1 - \alpha)^K \tag{3}$$

Then, for a given pixel in tampered blocks, the number of the usable subsets, n_U, obeys the following distribution:

$$P(n_U = t) = \binom{M}{t} \cdot (1 - p_U)^{M-t} \cdot p_U^t \tag{4}$$

For each usable subset, we check whether or not the extracted sum-bit is consistent with the modulus-2 sum

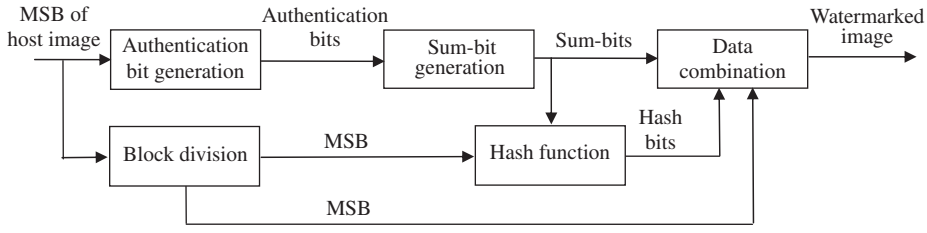

Fig. 1. Watermark embedding procedure.

of the pixel's authentication bit and other $(K-1)$ authentication bits. If, and only if, the consistency is satisfied in all usable subsets, the pixel is judged as "not tampered", indicating that there is no alteration in its 5 MSBs. Otherwise, it is judged as a "tampered" pixel. This way, a pixel without any alteration in its 5 MSBs must be judged as "not tampered", and probability with which a pixel containing modified MSBs is falsely judged as "not tampered" is

$$p_E = \sum_{t=0}^{M} [P(n_U = t) \cdot 2^{-t}] \quad (5)$$

Fig. 2 gives the p_E curves with different values of α and M. The narrower the modified area, the smaller the value of p_E is. If M is 40, the proposed scheme can exactly locate the tampered pixels when $\alpha < 5\%$. The performance is significantly better than the method in [5], which is effective only when the ratio between the number of tampered pixels and the image size is less than 1.1%. Denoting the number of tampered pixels with false "not tampered" judgments as N_E, its average value is

$$E(N_E) = p_E \cdot N_T \quad (6)$$

where N_T is the number of actual tampered pixels.

After finding a "tampered" pixel, we can further recover its original MSBs. The number of possible patterns of 5 MSBs is 32. We attempt to use 31 other patterns different from the received pattern of the pixel to check consistency between the extracted sum-bit and the modulus-2 sum of the pattern's authentication bit and other $(K-1)$ authentication bits. When consistency is arrived in all usable subsets, the attempted pattern is regarded as the original MSBs. If more than one pattern satisfies the consistency condition in all usable subsets, restoration of original pattern will be failed since we do not know which one is the true original pattern. The true original pattern must satisfy the consistency condition in all usable subsets, and the probability of the other pattern satisfies the consistency condition in all usable subsets is 2^{-n_U}. So, probability of failure is $1-(1-2^{-n_U})^{30}$. Considering the distribution of n_U, probability for the original MSBs of a pixel with "tampered" judgment being unable to find is

$$p_C = \sum_{t=0}^{M} \{P(n_U = t) \cdot [1-(1-2^{-t})^{30}]\} \quad (7)$$

Denoting the number of tampered pixels labeled "tampered" but being unable to be restored as N_C, its average is

$$E(N_C) = p_C \cdot (1-p_E) \cdot N_T \quad (8)$$

If $N_C = 0$, the receiver can obtain the original MSBs of all pixels, leading to restoration of the original watermarked image without any error.

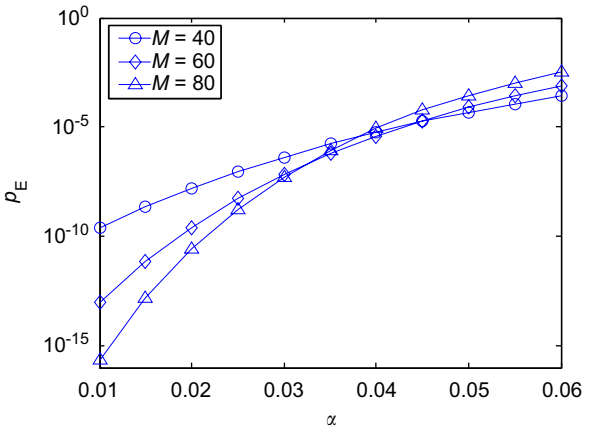

Fig. 2. Probabilities of false judgments with different values of α and M.

4. Discussion

As mentioned above, the proposed scheme is a combination of both block-wise and pixel-wise techniques. If the hash of the 320 MSBs and 160 sum-bits in a block is not identical to the 32 hash-bits hidden in the same block, the block is judged as "tampered". So, any modification on a block of watermarked image will result in a "tampered" decision with a probability $1-2^{-32}$. That means the watermark for identifying tampered blocks is completely fragile. When a watermarked image is compressed or low-pass filtered, most blocks are judged as "tampered", implying a large α. In this case, localization and restoration of tampered pixels will not be possible.

Consider a special case in which the tampered pixels are isolated and scattered over the image. As an extreme example, assume that each block in one-half of the watermarked image contains one tampered pixel. In other words, the proportion of tampered pixels is 1/128 and that of tampered blocks α is 1/2. Using the method in [5], all the tampered pixels can be localized, since the method is purely a pixel-wise approach, which is effective with a percentage of tampered pixels less than 1.1%. On the other hand, the present scheme will show that one half of the image is tampered but cannot locate the tampered pixels since the percentage of tampered blocks α is significantly more than 5%. In such case, the method in [5] is better. However, for malicious attacks that replace patches of image contents and the tampered pixels are not isolated, the scheme proposed in the present paper is better since it is effective with a higher percentage of tampered pixels/blocks and capable of recovering the original watermarked version.

5. Experimental results

Using a test image Lena sized 512×512 as the host and $M = 40$, the value of PSNR caused by watermark embedding is 37.9 dB, which verifies the theoretical analysis in [5] and is imperceptible. Fig. 3(a) and (b) show the watermarked image and a tampered version, in which four flowers were planted on the girl's hat with $N_T = 7550$ and $\alpha = 3.78\%$. By using the localization procedure, all the tampered pixels were found and shown in Fig. 3(c). Note that the MSBs of some pixels in the flowers coincided with the MSBs of original content, so these pixels were regarded as "not tampered" pixels. Furthermore, we can also recover the original MSBs of all tampered pixels using the restoration

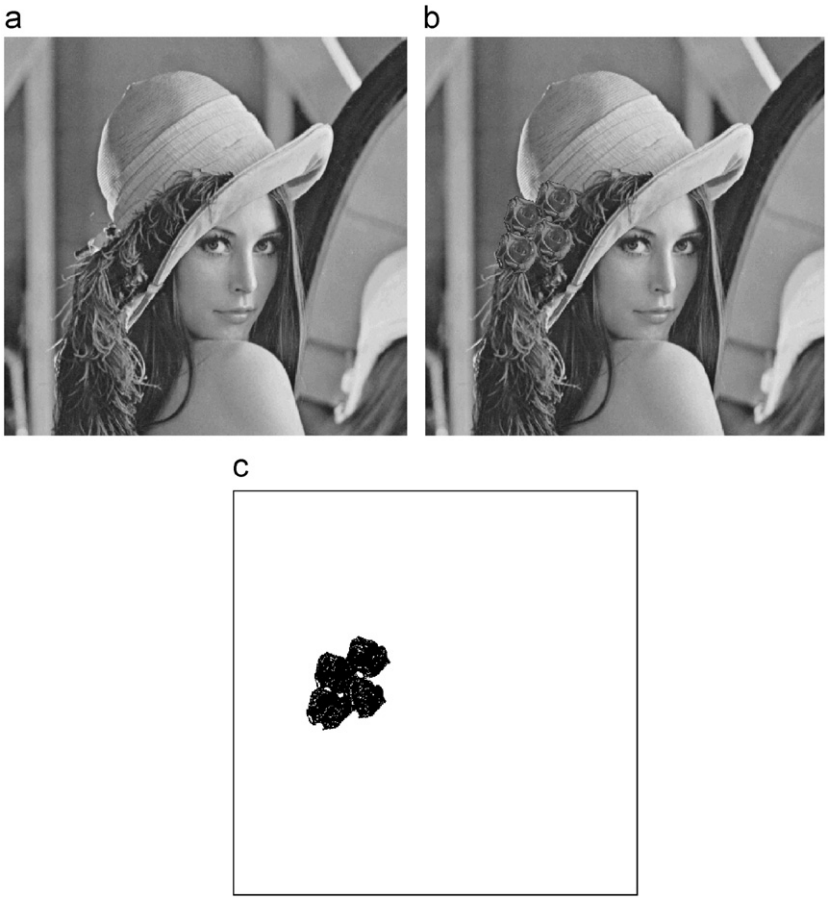

Fig. 3. (a) Watermarked Lena, (b) tampered version and (c) the tampered-pixel localization result.

Table 1
Theoretical and true values of N_E and N_C with different tampering strengths

N_T	6272	8185	10 015	11 910	13 844
α	3.30%	4.17%	5.05%	5.93%	6.59%
Theoretical value of N_E	0.01	0.07	0.53	2.88	8.86
True value of N_E	0	0	0	0	4
Theoretical value of N_C	0.18	2.12	15.54	79.95	231.65
True value of N_C	0	0	17	91	244

procedure. In other words, $N_C = 0$ and the original watermarked image can be perfectly restored on the receiver side.

In the experiment, we also replaced a portion of pixels in the watermarked image with different strengths, and then tried to locate the tampered pixels and to recover the original content. Table 1 lists the theoretical values of N_E and N_C in (6) and (8), and the corresponding true values. It can be seen that the theoretical and true values are very close. Experiments using other host images provided similar results.

6. Conclusion

The proposed fragile watermarking scheme based on a hierarchical mechanism is suitable for uncompressed host images, in which every pixel is represented by 8 bits. In this scheme, the watermark data derived from the MSBs are used to directly replace all the LSBs of a host image. Since the watermark embedding procedure works only in spatial domain, it is easy to implement. On the receiver side, after identifying the tampered blocks, the watermark hidden in the rest blocks are used to exactly locate the tampered pixels and to perfectly restore the original watermarked version. In comparison with the method described in [5], the proposed scheme possesses three advantages: (i) all tampered pixels can still be found when the modification area is significantly more extensive; (ii) it does not judge a pixel without any alteration in its five MSBs as "tampered", in other words, the false-positive probability is always zero, and (iii) the present scheme is capable of roughly recovering the original content and regenerating the watermarked version on the receiver side. In the future, we will develop fragile watermarking schemes based on transform domain techniques for images in a compression format such as JPEG.

Acknowledgments

This work was supported by the Natural Science Foundation of China (60502039, 60773079), the High-Tech

Research and Development Program of China (2007AA01Z477), and the Shanghai Leading Academic Discipline Project (T0102).

References

[1] P.W. Wong, N. Memon, Secret and public key image watermarking schemes for image authentication and ownership verification, IEEE Trans. Image Process. 10 (2001) 1593–1601.

[2] S. Suthaharan, Fragile image watermarking using a gradient image for improved localization and security, Pattern Recognition Lett. 25 (2004) 1893–1903.

[3] H. Lu, R. Shen, F.-L. Chung, Fragile watermarking scheme for image authentication, Electronics Lett. 39 (12) (2003) 898–900.

[4] S.-H. Liu, H.-X. Yao, W. Gao, Y.-L. Liu, An image fragile watermark scheme based on chaotic image pattern and pixel-pairs, Appl. Math. Comput. 185 (2) (2007) 869–882.

[5] X. Zhang, S. Wang, Statistical fragile watermarking capable of locating individual tampered pixels, IEEE Signal Process. Lett. 14 (10) (2007) 727–730.