Digital Watermarking Algorithm for Tampering Localization

Q. Chen, P.P. Xie and Y.Z. Hao

ABSTRACT

For image authentication, the competence of accurately locating tampering and ensuring security is desirable. In this paper, a novel reversible data embedding based on Integer Wavelet Transform for tamper-proof watermarks is proposed. A content-based watermark generated from host image is embedded into the carrier. The embedded data can be retrieved and the original host image can be recovered without distortion. The difference between the extracted watermark and the watermark generated from specific coefficients representing local features is applied to localize tampers. Experimental results show that the host image can be recovered after extracting the watermark if there are no attacks. Otherwise, the tampered positions can be localized explicitly and directly in the reconstructed host image even if the watermarked image was attacked by an invisible tampering. The proposed algorithm can be used for image authentication and tampering localization without using original image and original watermark.

KEYWORD: Image authentication; reversible watermarking; IWT; tampering localization

INTRODUCTION

Image authentication based on digital watermarking is an important part of the image authentication. But most of the current methods will introduce certain distortions in the process of embedding authentication information. For general applications, these distortions can be tolerated, but for the distortion-sensitive applications such as medical images, military image, or evidence to be identified in court, there is no room for any distortion.

Q. Chen, P.P. Xie, Y.Z. Hao. University of Shanghai for Science and Technology, Shanghai, China.
Reversible watermarking is the process that allows extraction of embedded information and exact recovery of the original image from the watermarked image which does not changed during transmission [1-4].

Tian [1] introduced an algorithm for reversible watermarking using different expansion. The scheme embeds the information into the least significant bits (LSB) of the expanded differences of pixel pairs. Lee et al. [5] proposed a scheme in which an input image is divided into blocks, and a watermark is embedded into the high-frequency wavelet coefficients of each block by LSB-substitution or bit-shifting. The original image can be exactly reconstructed at the decoder since the side information for achieving the reversibility is also embedded in the image while avoiding the underflow and overflow. In the algorithm [6], the position information with small size is kept as a secret matrix. However, those [1,4-6] are semi-blind. The overhead information like location map, position or side information are embedded or needed; what is more, all of them are not capable to localize the tampered area.

An invertible authentication watermarking [7] is proposed for enabling zero-distortion reconstruction of the un-watermarked images upon verification. For verified images, integrity of the reconstructed image is ensured by the uniqueness of the reconstruction procedure. The framework also enables public authentication without granting access to the perfect original and allows for efficient tamper localization.

In response to the deficiencies, an IWT-based reversible blind watermarking with characteristic of tampering localization is proposed. The algorithm can not only reconstruct carrier images without distortion, but also localize the tampered area using the differences between the extracted watermark and authentication watermark constructed via the content of carrier image, without the aid of the original watermark or host image even if watermarked image might be attacked by an invisible tampering. The experimental results demonstrate its efficiency.

**AUTHENTICATION BY WATERMARKING**

Watermarking-based image authentication [8,9] generates authentication watermark associated with the original image and embeds the watermark into the original image in an imperceptible way, so that it is difficult for the attackers to detect the presence of authentication information or to extract it.

![Figure 1. Block diagram of watermarking-based authentication.](a)  
![Figure 1. Block diagram of watermarking-based authentication.](b)
As shown in Fig 1, the authentication watermark, denoted as $M_1$, can be obtained from the local features of a host image and embedded into the host image. At the receiving end, authenticator obtains reference watermark, denoted as $M_2$, from the received carrier image $Y$ by using the same watermark-generating method. In the meanwhile, authenticator recovers the embedding watermark, denoted as $M'_1$, from $Y$ by watermark extracting algorithm. Image is authenticated by comparing the reference watermark $M_2$ and the extracted watermark $M'_1$. The rule is if $M'_1 \approx M_2$ it means $M'_1$ is much similar or equivalent to $M_2$; Else if $M'_1 \neq M_2$, which means image content has changed. Because the construction and embedding of watermark have close relation with local image feature, $M'_1 \neq M_2$ means that the local features used to generate and embed watermark may have changed, so that pixels related to the local features and the characteristic coefficients may also have changed.

Construction of Authentication Watermark

Original host is decomposed by IWT into different sub-bands after a 3-level IWT. The sub-band $HH_3$ is selected to generate authentication watermark $M_1$. Because the edge features of image, maintaining their position in the space, are in $HH_3$, they guarantee the invariance of main content of the formed watermark. The two-value authentication watermark $M_1$ is generated by binarizing the coefficients of sub-band $HH_3$ with threshold value $T$. Then this watermark information is embedded into sub-band $LH_2$ by adjusting the selected coefficients according to watermark embedding algorithm.

Watermark Embedding and Extraction

Suppose that the size of sub-band to be embedded watermark is $m \times m$ and the size of watermark is $n \times n$. Here, $m = 2n$. First, the coefficients of sub-band to be embedded are divided into blocks, so that one bit of watermark is embedded into one block. Take a $2 \times 2$ block (denoted as A) as an example. It consists of 4 coefficients denoted by $a_i$ ($1 \leq i \leq 4$). It is noticeable that these 4 coefficients belong to integer due to integer wavelet transform. One bit of watermark (denoted as $b$ ($b \in \{0,1\}$)) is embedded into the coefficient with minimum absolute value (denoted as $a_{min}$) among coefficients in block A.

$$a_{min} = \min(|a_i|) \quad (1 \leq i \leq 4)$$  \hspace{1cm} (1)

Watermark embedding is as follows.

$$\begin{align*}
\hat{a}_i &= a_i + (a_{min} + 1), \quad \text{if} \quad a_i > a_{min} \\
\hat{a}_i &= 2 \times a_i + b, \quad \text{if} \quad a_i = a_{min} \\
\hat{a}_i &= 2 \times a_i - b, \quad \text{if} \quad a_i = -a_{min} \\
\hat{a}_i &= a_i - (a_{min} + 1), \quad \text{if} \quad a_i < -a_{min}
\end{align*}$$

(2)

Watermark extraction is the inverse procedure of embedding. Firstly, find out the coefficient with the minimum absolute value $\hat{a}_{min} = \min(|\hat{a}_i|)$.

$$b = \text{mod}(\hat{a}_{min}, 2)$$  \hspace{1cm} (3)
Here "mod (.)" represents modulo operation. From the above formulas, the coefficients which were changed in the process of watermark embedding can be recovered to the original value $a_i$.

**Tampering Localization**

If the watermarked host image suffers imperceptible tampering, the tampered area in the watermarked host image can still be directly revealed and even intuitively localized by positioning watermark. The positioning watermark $\hat{M}$, which is difference image between $M_2$ and $M_1$, is used to label the corresponding coefficients in low-frequency sub-band, in addition to show the tampered areas in the positioning watermark for comparison. We can seek out the coefficients corresponding to positioning watermark and replace their values with zero as follows. Then, we can get the reconstructed image with tampering localization after an inverse IWT.

$$
\begin{aligned}
    d_{ij} &= 0, \quad \text{if} \quad \hat{m}_{ij} = 1 \\
    d_{ij} &= d_{ij}, \quad \text{if} \quad \hat{m}_{ij} = 0
\end{aligned}
$$

(5)

Where $\hat{m}_{ij}$ represents $(i,j)$ point in positioning watermark $\hat{M}$, $d_{ij}$ represents the $(i,j)$ $(1 \leq i,j \leq n)$ coefficient in sub-band $LL_3$, and $n \times n$ its size.

**EXPERIMENTS AND ANALYSIS**

**Reliability of reference watermark.** The reference watermark $M_2$ is generated from the received watermarked host image. Reliability of reference watermark means $M_2$ must be much similar or identical to authentication watermark $M_1$. Because of the effects of watermark embedding and communication transmission, it is not guaranteed that $M_2$ still share the same characteristic of $M_1$. Therefore, reliability of $M_2$ is estimated by NC (normalized correlation) between $M_2$ and $M_1$.

As is shown in Table 1, even the watermarked host image has been under attack, $M_2$ and $M_1$ still have a high similarity in the proposed algorithm when the threshold $T$ is larger than 20. We can come to the conclusion that stability can be ensured if the watermark-generating threshold $T$ is given properly. Hence, reference watermark $M_2$ constructed from attacked host image remain quite reliable for image authentication and tampering localization without using original authentication watermark $M_1$. 
Table 1. Similarity (NC) between M2 and M1 with different thresholds when embedded image is under various attacks.

| NC | speckle noise intensity factor \( \sigma^2 = 0.002 \) | speckle noise intensity factor \( \sigma^2 = 0.05 \) | salt & pepper noise | cropping area \( 1/25 \) | JPEG compression factor \( 20 \) | average filtering window size \( 5 \times 5 \) |
|----|---------------|---------------|----------------|----------------|----------------|----------------|
| NC (T=40) | 0.9978 | 0.9766 | 0.9726 | 0.9998 | 0.9973 | 0.9995 |
| NC (T=20) | 0.9892 | 0.8111 | 0.8200 | 0.9995 | 0.9921 | 0.9974 |
| NC (T=3) | 0.8307 | 0.6320 | 0.6545 | 0.9974 | 0.9267 | 0.9655 |

Figure 2. Sensitivity experiment (a) watermarked carrier attacked by speckle noise \( \sigma^2 = 0.002 \); (b) reference watermark \( M_2 \); (c) extracted watermark \( M'_1 \).

**Sensitivity experiment of watermark.** Watermark generating threshold \( T \) is set to 40 and similarity \( (NC) \) is used to judge the authenticity of watermark. The \( NC \) values between original authentication watermark and 200 random watermarks are not greater than 0.5300, therefore, if similarity between \( M_2 \) (reference watermark) and \( M'_1 \) (extracted watermark) is greater than 0.55 (watermark-detecting threshold \( \tau \)), the extracted watermark can be considered as valid so the host image should be credible. Otherwise, the watermark can be considered as invalid that means the host image suffers from tamper or attack. After speckle noise attack \( \sigma^2 = 0.002 \) simulated to the watermarked image, experimental result (Fig. 2) shows that the similarity \( NC \) between \( M_2 \) and \( M'_1 \) is equal to 0.5105, less than the threshold \( \tau = 0.55 \), demonstrating that the content of the image is incredible. As Table 2 shown, although low-intensity attack is imperceptible to watermarked carrier, the embedded watermark \( M_1 \) has been severely damaged \( (NC < \tau) \) indicating that watermark is fragile and very sensitive to slight attack.
Table 2. Similarity (NC) between $M'_1$ and $M_2$ under various attacks.

| attack  | parameter | PSNR  | NC   |
|---------|-----------|-------|------|
| speckle | 0.002     | 32.1151 | 0.5105 |
| noise   | 0.005     | 28.4367 | 0.5030 |
|         | 0.01      | 25.5392 | 0.4963 |
|         | 0.02      | 22.6103 | 0.4943 |
|         | 0.05      | 18.8005 | 0.4910 |
| pepper  | 0.002     | 32.0360 | 0.5135 |
| &       | 0.005     | 28.2573 | 0.5110 |
| salt/noise | 0.01   | 25.2170 | 0.5037 |
| e       | 0.02      | 22.4040 | 0.4952 |
|         | 0.05      | 18.3359 | 0.4903 |
| median  | 2×2       | 28.1726 | 0.4978 |
| filtering | 3×3     | 33.7552 | 0.4930 |
|         | 4×4       | 27.6892 | 0.5070 |
|         | 5×5       | 29.7190 | 0.5002 |
|         | 7×7       | 27.7060 | 0.5107 |
| JPEG   | 90%       | 36.8096 | 0.5227 |
| compress | 75%      | 35.0287 | 0.5030 |
| ion    | 60%       | 34.2665 | 0.4815 |
|         | 45%       | 33.4072 | 0.4484 |
|         | 20%       | 31.2863 | 0.3459 |

**Test on transparency of host image.** Threshold $T$ is still set to 40. From the experimental results, equivalence between the original images and reconstructed images has proved the reversibility of the proposed algorithm. The comparison of the original host image and the watermarked image without any attack is shown in Fig 3(a) and (b) respectively. The visual qualities of the watermarked image are satisfactory. There is no obvious visual distortion in watermarked image and its PSNR is 44.0062. The similarity (NC) between extracted watermark and the original one is 1, which means that watermark can be losslessly extracted.

![Figure 3. Transparency test: (a) original host; (b) watermarked image (PSNR=44.0062); (c) extracted watermark (NC=1).](image)

**Experiments on tampering localization.** In following test, the watermarked image suffered tamper attacks, and the star is erased from the tail of plane in the host image (Fig 4). It is evident that tampered area is not noticeable at the receiving end if there is no original host image by contrast. The proposed algorithm makes it possible that tampered area can be positioned directly in the reconstructed host image despite the lack of original host image. Corresponding extracted watermark $M_1$, reference watermark $M_2$; and positioning watermark $\hat{M}$ are shown in Fig. 5. The tampered blocks can be indicated clearly in the reconstructed host image (Fig 4 (c)) with the help of the positioning watermark $\hat{M}$ ($\hat{M} = M_2 - M'_1$) (Fig 5 (c)). It is obvious that the unsmooth and irregular black chunk on the tail of plane (Fig 4 (c)) indicates the tampered area.
Figure 4. Test 2 on tampering localization: (a) Original host image; (b) watermarked image after being tampered; (c) reconstructed host image after tampering localization.

Figure 5. (a) Extracted watermark $M_1$; (b) Reference watermark $M_2$; (c) Positioning watermark $\hat{M}$.

CONCLUSION

The paper proposed an IWT-based reversible blind watermarking method. In this scheme, a content-based watermark is embedded into the host image for image authentication and tampering localization. Experimental results demonstrate that the host image can be losslessly recovered from the watermarked host image if there are no attacks. Otherwise, the tampered positions can be localized explicitly and directly in the reconstructed host image without the aid of original image and original watermark, even if the watermarked image was attacked by an invisible tampering. Since it is built on discrete wavelet transformation, it can be integrated with JPEG2000 in the near future. Further work is to build new methods to locate the tampered blocks more precisely and approximately recover the tampered region.

ACKNOWLEDGEMENTS

The work on this paper was carried out partially under the financial support by the State 863 projects (No.2012AA050206), Shanghai Natural Science Foundation (No.12ZR1420800), and National Project Cultivating Fund of University of Shanghai for Science and Technology (16HJPY-MS06).

REFERENCES

1. J. Tian, “Reversible data embedding using a difference expansion,” IEEE Trans. Circuits Syst. Video Technol., vol. 13, no. 8, p 890-896, Aug. 2003.
2. Zhicheng Ni, Yun-Qing Shi, Nirwan Ansari, and Wei Su, “Reversible data hiding,” IEEE Transactions on circuits and systems for video technology, vol. 16, no. 3, p354-362, march 2006.
3. Lixin Luo, Zhenyong Chen, Ming Chen, Xiao Zeng, and Zhang Xiong “Reversible image watermarking using interpolation technique,” IEEE Transactions on information forensics and security, vol. 5, no. 1, p187-193, march. 2010.
4. Shu-Kei Yip, Oscar C. Au, Chi-Wang Ho, Hoi-Ming Wong , “Lossless Visible Watermarking,” IEEE International Conference on Multimedia and Expo, p 853-856, July 2006.
5. Sunil Lee, Chang D. Yoo, and Ton Kalker, “Reversible image watermarking based on Integer-to-integer wavelet transforms,” IEEE Transactions on information forensics and security, vol. 2, no. 3, p 321-330, September 2007.
6. QiaoLun Gu, Guanglin Han, Tiegang Gao, Zengqiang Chen, “A Novel Adaptive Reversible Watermarking Algorithm Based on Wavelet Lifting Scheme,” IEEE International Conference on Information Engineering and Computer Science, 2009. p1-4, Dec. 2009.

7. Mehmet Utku Celik, Gaurav Sharma, Murat Tekalp, “Lossless Watermarking for Image Authentication: A New Framework and an Implementation”, IEEE TRANSACTIONS ON IMAGE PROCESSING, VOL. 15, NO. 4, APRIL 2006.

8. Phadikar A., Maity S.P., Mandal M. “QIM data hiding for tamper detection and correction in digital images using wavelet transform,” 2010 23rd Canadian Conference on Electrical and Computer Engineering, p 1-5, 2010.

9. Mingju Chen, Xingbo Sun, “A digital image watermarking of self-recovery base on the SPIHT algorithm,” IEEE 2nd International Conference on Signal Processing Systems, p 621-624, 2010.