J. Vis. Commun. Image R. 30 (2015) 312–327

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

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

A singular-value-based semi-fragile watermarking scheme for image content authentication with tamper localization ☆

Xiaojun Qi *, Xing Xin

Department of Computer Science, Utah State University, Logan, UT 84322-4205, United States

ARTICLE INFO

Article history:
Received 10 July 2014
Accepted 13 May 2015
Available online 21 May 2015

Keywords:
Adaptive quantization
Authentication measures
Tampered error map
Semi-fragile watermarking
Singular value decomposition
Wavelet transform
Content-dependent watermark
Strongly tampered error map
Mildly tampered error map

ABSTRACT

This paper presents a singular-value-based semi-fragile watermarking scheme for image content authentication. The proposed scheme generates secure watermark by performing a logical operation on content-dependent watermark generated by a singular-value-based sequence and content-independent watermark generated by a private-key-based sequence. It next employs the adaptive quantization method to embed secure watermark in approximation subband of each 4 × 4 block to generate the watermarked image. The watermark extraction process then extracts watermark using the parity of quantization results from the probe image. The authentication process starts with regenerating secure watermark following the same process. It then constructs error maps to compute five authentication measures and performs a three-level process to authenticate image content and localize tampered areas. Extensive experimental results show that the proposed scheme outperforms five peer schemes and its two variant systems and is capable of identifying intentional tampering, incidental modification, and localizing tampered regions under mild to severe content-preserving modifications.

1. Introduction and related work

With the amount of digital multimedia documents (image, text, audio, and video) growing at an accelerating rate, the development of effective techniques to reinforce their security has become an active research area in recent years [1]. Especially, digital multimedia authentication has become an important issue since multimedia documents can be easily duplicated, modified, transformed, and diffused with the aid of digital multimedia editing software and trustworthy multimedia plays a vital role in applications including news reporting, intelligence information gathering, criminal investigation, security surveillance, and health care. This paper aims to develop an effective technique for digital multimedia authentication.

Recently, fragile, semi-fragile, and self-recovering watermarking techniques have been considered as potential promising techniques for multimedia authentication. Fragile watermarking techniques aim to be fragile to any modifications so as to detect and localize these modifications. Semi-fragile watermarking techniques aim to resist acceptable content-preserving modifications and detect malicious content-altering modifications. Self-recovering watermarking techniques incorporate the content recovery property in either fragile or semi-fragile schemes to not only detect and localize the modifications but also recover the original content. However, self-recovering fragile techniques work well when the tampered area is not extensive [2–5] and self-recovering semi-fragile techniques work well under mild or no content-preserving modifications [6–9]. In this paper, we propose a singular-value-based (SV-based) semi-fragile watermarking technique that is capable of detecting various small to large content modifications and localizing tampered areas even under mild to severe content-preserving modifications.

Semi-fragile watermarking algorithms can be classified into spatial domain-based and transform domain-based schemes [10]. In general, spatial domain-based schemes embed watermark into the host image by modifying a set of pixel values without causing obvious changes in its appearance. Transform domain-based schemes embed watermark into the host image by modifying transformed coefficients such as discrete cosine transform (DCT) coefficients and discrete wavelet transform (DWT) coefficients. Both schemes use extracted watermark to authenticate the digital content and localize tampered areas if possible. Transform domain-based schemes are better than spatial domain-based schemes since they tend to achieve better invisibility and more robustness. Here, we briefly review several representative

☆ This paper has been recommended for acceptance by Prof. M.T. Sun.

* Corresponding author. Fax: +1 435 7973265.
E-mail address: Xiaojun.Qi@usu.edu (X. Qi).

http://dx.doi.org/10.1016/j.jvcir.2015.05.006
1047-3203/© 2015 Elsevier Inc. All rights reserved.

semi-fragile watermarking schemes in two popular transform domains (DCT and DWT domains).

Ho and Li [11] use the relationship of DCT coefficients in low and middle frequencies to protect the authenticity of a compressed watermarked image when the JPEG quality is higher than authors' predefined lowest authenticable quality. Maeno et al. [12] propose two semi-fragile authentication techniques. The first one incorporates a random bias factor to the fixed decision boundary to detect malicious manipulations and keep the false alarm rate low. The second one uses a non-uniform quantization scheme to improve the encoding accuracy of relationships between paired DCT coefficients and achieve higher alteration detection sensitivity. Zhou et al. [13] propose to extract a signature from non-overlapping blocks of the original image and insert it into selected block-based wavelet coefficients. Hu and Han [14] use the hash function to generate one watermark for classifying the intentional content modification and use the Sobel edge detector to generate the other watermark for indicating the modification location. These two watermarks are finally embedded in the middle-frequency wavelet coefficients. Zhu et al. [15] apply the block-mean quantization strategy to embed inter-block and intra-block signatures in the finest scale of wavelet coefficients for tamper detection and localization, respectively. Yang and Sun [16] integrate the human visual system (HVS) model in the embedding scheme to insert watermark by modifying vertical and horizontal subbands of image sub-blocks. Two measures are then used to judge whether a modification is malicious. Che et al. [17] propose the HVS-based dynamic quantization approach to embed watermark in low-frequency wavelet coefficients. Cruz et al. [18] propose to extract a robust signature from 16×16 non-overlapping blocks by thresholding projections onto random smooth patterns. They then employ the vector quantization method [13] to embed this signature into the approximation subband of each image sub-block. Zhang et al. [19] design a fast scheme to avoid the overflow checking of pixel intensity by adjusting wavelet coefficients and adopt the cumulative weighted voting method to reduce the false alarm pixels. Preda [20] propose to embed a watermark bit in a group of randomly permuted wavelet coefficients by means of quantization. The embedded watermark is robust to mild to moderate JPEG compression by selecting appropriate embedding parameters. Morphological operations are also used to improve detection results. Huo et al. [21] propose to embed general tampering and collage attack watermarks in the block-based DCT domain and detect tampered regions based on the two maps generated from two extracted watermarks.

However, all these schemes are only robust to moderate JPEG compression of higher than a 50% or 60% quality factor (QF). The false alarm rates for watermarking schemes proposed in [13–15,18,20] are high under common image processing attacks. To address these shortcomings, two improved schemes have been recently proposed. Qi and Xin [22] employ a non-traditional quantization method to embed watermark by modifying one chosen DWT approximation coefficient of each non-overlapping block. They then compute two measures to confirm the image content and localize possible tampered areas. Al-Otum [23] embeds a random watermark bit sequence into the DWT domain using an adjusted expanded-bit multi-scale quantization index modulation (QIM) [24] based technique. Two measures are further developed to classify the probe image as authenticated, incidentally, or maliciously attacked. However, the two measures in both schemes may fail to recognize tampered blocks if the watermarked image is also incidentally attacked by a moderate noise signal or a higher JPEG or JPEG2000 compression. Furthermore, the technique proposed in [23] is limited in its use due to its inability to process color images as mentioned in the conclusion section.

In this paper, we propose an effective semi-fragile watermarking scheme that is capable of detecting various small to large content modifications and localizing tampered areas even under mild to severe content-preserving modifications. It first generates secure watermark by performing the logical "*xor*" operation on content-dependent and content-independent watermarks. Here, content-dependent watermark is a robust signature extracted by SV-based features and content-independent watermark is a private-key-based random watermark. The proposed scheme then uses the adaptive quantization method to embed secure watermark in the wavelet domain. It further utilizes a three-level authentication process to authenticate the image content and localize tampered areas. Specifically, the first-level process employs two measures (e.g., M_1 and M_2) derived from a binary error map to quickly classify the maliciously attacked watermarked image under mild content-preserving modifications. The second-level process employs another two measures (e.g., M_3 and M_4) derived from a binary strongly tampered error map to further classify the probe image as authenticated, incidentally, or maliciously attacked under moderate to severe content-preserving modifications. The third-level process employs the fifth measure (e.g., M_5) derived from the post-processed candidate tampered area to further validate the nature of the attack and produce a clean tampered result if applicable. This scheme also possesses all the desired properties (e.g., invisibility, tamper detection, security, identification of tampered areas, oblivion with no transmission of any secret information, and discrimination of incidental distortion and malicious tampering) for an effective authentication watermarking scheme [25]. Our contributions are:

- Utilizing relationships of SVs among three sub-blocks of each 4×4 non-overlapping JPEG quantized block to extract content-dependent watermark for both watermark embedding and authentication processes.
- Generating secure watermark by performing the logical "*xor*" operation on SV-based content-dependent watermark and private-key-based content-independent watermark.
- Merging relationships of SVs among three sub-blocks of each 4×4 non-overlapping JPEG quantized block to choose its adaptive quantizer q for both watermark embedding and extraction processes.
- Applying the adaptive quantization method to embed secure watermark in the wavelet domain so that a majority of image distortions, which cause the intensity shift by a value larger than a half of the quantizer q or cause the image content to change, can be detected in the authentication process.
- Defining a three-level authentication process involving five measures to quantitatively detect the authenticity of the probe image and prove tampering, with M_1, M_2, M_3, M_4, and M_5 measuring the similarity between extracted and embedded watermarks, the clustering level of tampered error pixels, the average size of connected components formed by strongly tampered error pixels, the variation in the sizes of connected components, and the clustering level of tampered error pixels around the potential distorted area, respectively.
- Using five authentication measures derived from three binary maps (e.g., error map, strongly tampered error map, and mildly tampered error map) to compensate possible misclassification and ensure that the proposed scheme is able to capture distortions and localize tampered areas when the probe image is also incidentally attacked by a moderate noise signal or a higher compression.

The remainder of the paper is organized as follows: Section 2 presents the proposed semi-fragile watermarking scheme. Section 3 quantitatively evaluates the performance of the proposed scheme. Section 4 demonstrates the effectiveness of the proposed scheme by comparing it with its two variant schemes and five peer schemes [12,16–18,20] on extensive experiments. Section 5 draws the conclusion.

2. The proposed scheme

The proposed scheme consists of four components: secure watermark generation, watermark embedding, watermark extraction, and watermark authentication. In the following subsections, we explain each component in detail.

2.1. Secure watermark generation

Using relationships of SVs among three sub-blocks of each 4 × 4 non-overlapping JPEG quantized block of the original image, we generate content-dependent watermark *CW* that represents intrinsic algebraic image properties to facilitate the authentication process. In order to increase the security, we also generate random content-independent watermark *IW* by using the Mersenne Twister algorithm [26] and a private key k_1. Secure watermark *SW* is finally generated by performing the logic "*xor*" operation on *CW* and *IW*. This secure watermark encodes the image content and therefore is more fragile to content altering operations and moderate to severe common image processing attacks. The flowchart of secure watermark generation is shown in Fig. 1.

The quantized image *q-I* is generated by modifying coefficients (i.e., $Blk_i(x,y)$) in each 8 × 8 block Blk_i of original image *I* to an integral multiple (i.e., $modified\_Blk_i(x,y)$) of the quantization matrix *Q* as follows:

$$modified\_Blk_i(x,y) = round(Blk_i(x,y)/Q(x,y)) \times Q(x,y) \qquad (1)$$

where *round*( ) represents the rounding operation, *Q* is the quantization matrix specified in the JPEG standard ($1 \leqslant x, y \leqslant 8$), and operations / and × are the element-wise division and multiplication, respectively.

For each 4 × 4 quantized block q_Blk_i, the following operations are performed:

1. Divide the block into 2 × 2 sub-blocks to obtain $subblock_{i\_0}$, $subblock_{i\_1}$, $subblock_{i\_2}$, and $subblock_{i\_3}$ in the raster scan order.
2. Apply SVD on $subblock_{i\_1}$ to obtain three matrices U_{i_1}, S_{i_1}, and V_{i_1}, where $subblock_{i\_1} = U_{i\_1} \times S_{i\_1} \times V^T_{i\_1}$. Similarly, apply SVD on $subblock_{i\_2}$ and $subblock_{i\_3}$ to obtain two sets of three matrices, i.e., U_{i_2}, S_{i_2}, and V_{i_2}, and U_{i_3}, S_{i_3}, and V_{i_3}, respectively.
3. Generate a watermark bit based on the relationship of the most notable and stable SVs of $subblock_{i\_1}$, $subblock_{i\_2}$, and $subblock_{i\_3}$, which correspond to three values, i.e., $S_{i\_1}(1,1)$, $S_{i\_2}(1,1)$, and $S_{i\_3}(1,1)$. The rules for generating content-dependent watermark bit CW_i are:
   3.1 Generate bit B_1 using the relationship between $S_{i\_1}(1,1)$ and $S_{i\_2}(1,1)$ by:

$$B_1 = \begin{cases} 1 & \text{if } S_{i\_1}(1,1) \geqslant S_{i\_2}(1,1) \\ 0 & \text{otherwise} \end{cases} \qquad (2)$$

   3.2. Similarly, generate bit B_2 using the relationship between $S_{i\_2}(1,1)$ and $S_{i\_3}(1,1)$ and generate bit B_3 using the relationship between $S_{i\_1}(1,1)$ and $S_{i\_3}(1,1)$, respectively.

   3.3 Generate CW_i by applying a series of "*xor*" operations:

$$CW_i = xor(xor(B_1, B_2), B_3) \qquad (3)$$

   3.4 Evaluate the intrinsic algebraic image property of the block by:

$$PBlock_i = B_1 + B_2 + B_3 \qquad (4)$$

We choose SVs to generate *CW* due to its stability in terms of a small perturbation, its invariant algebraic and geometric properties, and its compact representation of algebraic properties of an image [27]. The relationship of SVs between sub-block pairs measures the luminant difference of each pair. The $PBlock_i$ value for each quantized block measures the luminant change patterns of each pair, which is used to automatically determine the block's quantization step. The $subblock_{i\_0}$ is not used to generate the watermark. Instead, it is used for embedding watermark to ensure the robustness of generating *CW* and *SW*.

2.2. Watermark embedding

We divide the original image into non-overlapping 4 × 4 blocks and embed secure watermark *SW* in the wavelet domain of each unique randomly chosen 4 × 4 block. The flowchart of the watermark embedding process is shown in Fig. 2.

We choose the wavelet domain as the embedding media mainly due to its excellent spatial-frequency localization and its compatibility with the JPEG2000 coding standard. Specifically, we utilize the parity of the quantized value at the approximation subband to embed watermark. To ensure the watermark's invisibility and increase robustness against common image processing attacks, we choose the upper-left value of the approximation subband

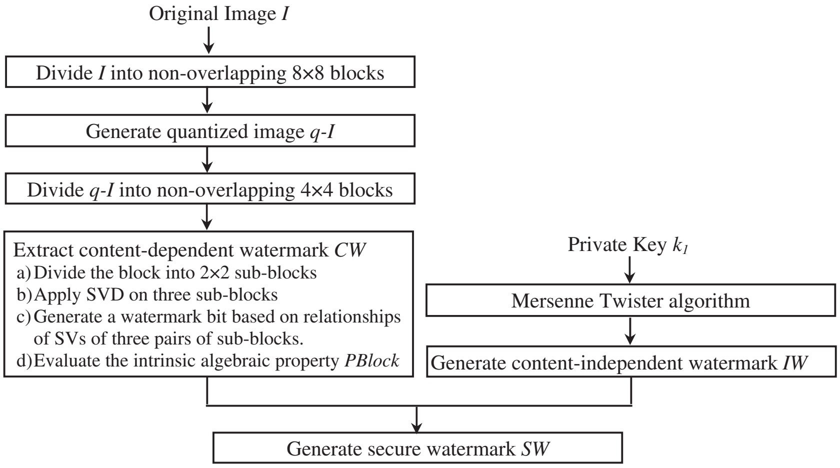

Original Image I
Divide I into non-overlapping 8×8 blocks
Generate quantized image q-I
Divide q-I into non-overlapping 4×4 blocks
Extract content-dependent watermark CW
a) Divide the block into 2×2 sub-blocks
b) Apply SVD on three sub-blocks
c) Generate a watermark bit based on relationships of SVs of three pairs of sub-blocks.
d) Evaluate the intrinsic algebraic property PBlock
Private Key k1
Mersenne Twister algorithm
Generate content-independent watermark IW
Generate secure watermark SW

Fig. 1. Flowchart of the secure watermark generation process.

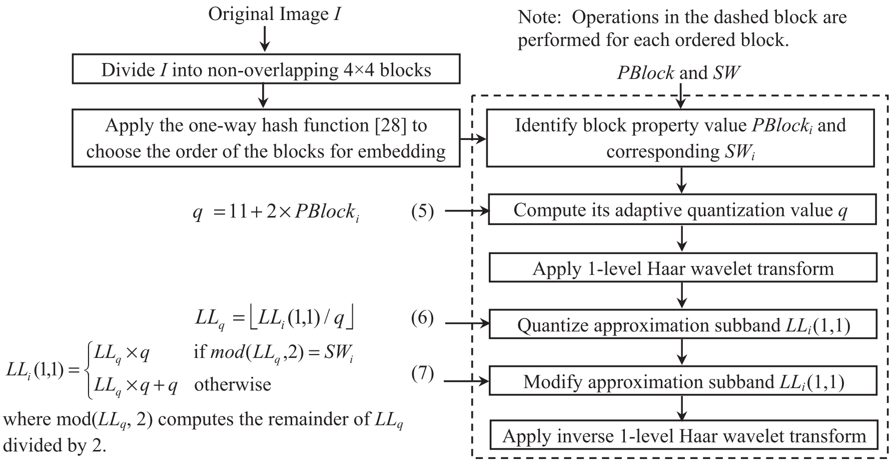

Original Image I
Divide I into non-overlapping 4×4 blocks
Apply the one-way hash function [28] to choose the order of the blocks for embedding
Note: Operations in the dashed block are performed for each ordered block.
PBlock and SW
Identify block property value $PBlock_i$ and corresponding SW_i
$q = 11 + 2 \times PBlock_i$ (5)
Compute its adaptive quantization value q
Apply 1-level Haar wavelet transform
$LL_q = \lfloor LL_i(1,1)/q \rfloor$ (6)
Quantize approximation subband $LL_i(1,1)$
$LL_i(1,1) = \begin{cases} LL_q \times q & \text{if } mod(LL_q,2) = SW_i \\ LL_q \times q + q & \text{otherwise} \end{cases}$ (7)
where mod(LL_q, 2) computes the remainder of LL_q divided by 2.
Modify approximation subband $LL_i(1,1)$
Apply inverse 1-level Haar wavelet transform

Fig. 2. Flowchart of the watermark embedding process. (See above-mentioned references for further information.)

(e.g., $LL_i(1,1)$), which corresponds to $subblock_{i\_0}$, to embed a watermark bit. The embedding strategy is to ensure that the parity of the modified $LL_i(1,1)$ is consistent with the embedding bit. It should be noted that the bigger the q is, the bigger the changes, consequently, the worse the quality of the watermarked image, and the stronger the robustness. In our system, q is adaptively chosen for each block B_i based on $PBlock_i$. A larger q value is assigned to a block with a higher $PBlock_i$ value. The value of q can be 11, 13, 15, or 17 for four $PBlock_i$ values (i.e., 0, 1, 2, and 3), respectively. These four quantization values are determined based on the tradeoff among invisibility, robustness, and fragileness. They have been shown to be effective when using either of them as the fixed quantizer in the proposed system.

2.3. Watermark extraction

The watermark extraction process is similar to the watermark embedding process except that it uses the parity of the quantized upper-left value of the approximation subband of each non-overlapping 4×4 block to extract the watermark bit. The upper-left value $LL(1,1)'$ of each block is quantized by the adaptive quantizer q calculated by Eq. (5) as follows:

$$LL'_q = \mathrm{round}(LL(1,1)'/q) \tag{8}$$

The watermarked bit EW_i for each block is extracted as mod(LL_q', 2).

2.4. Watermark authentication

First of all, we simulate the process of generating secure watermark SW' by applying the logical "*xor*" operation on regenerated content-dependent watermark CW' and content-independent watermark IW'.

Second, we perform the first-level authentication using measures M_1 and M_2 computed from the binary error map *ErrorMap*. This map is generated by mapping the absolute difference between extracted and regenerated secure watermarks (e.g., $|EW_i - SW_i'|$) onto its corresponding 4×4 block with 0's and 1's indicating match and mismatch, respectively. In other words, any pixel with the value of 1's is an error pixel. In the proposed system, we classify tampered error pixels in a 3×3 window in *ErrorMap* into two categories: strongly and mildly tampered error pixels. An error pixel is strongly tampered if at least five of its eight neighbors are error pixels; and an error pixel is mildly tampered if at most four of its eight neighbors are error pixels. We define two authentication measures by:

$$M_1 = \frac{\text{Number of error pixels in } \mathit{ErrorMap}}{\text{Number of pixels in } \mathit{ErrorMap}} \tag{9}$$

$$M_2 = \frac{\text{Number of strongly tampered error pixels in } \mathit{ErrorMap}}{\text{Number of tampered error pixels in } \mathit{ErrorMap}} \tag{10}$$

Here, M_1 measures the similarity between extracted and regenerated secure watermarks and M_2 measures the clustering level of tampered error pixels. M_2 is set to 0's if the count of tampered error pixels is zero. This authentication process aims to quickly identify tampered regions for the probe image undergoing content-altering modifications together with mild content-preserving modifications. We apply median filtering on *ErrorMap* when M_1 is less than or equal to T_{median} to accommodate the small amount of distortions caused by mild modifications. This filtering keeps clustered error pixels intact and makes scattered mildly tampered and isolated error pixels disappear. As a result, the small malicious attack leads to a larger M_2 value due to the removal of mildly distorted error pixels.

Third, we perform the second-level authentication using measures M_3 and M_4 computed from the binary error map *STErrorMap*, which is generated by marking strongly tampered error pixels in *ErrorMap* as 1's. We extract connected components in *STErrorMap* and define two other authentication measures as follows:

$$M_3 = \frac{\text{Number of error pixels in } \mathit{STErrorMap}}{\text{Number of connected components in } \mathit{STErrorMap}} \tag{11}$$

$$M_4 = std(\text{the sizes of all connected components in } \mathit{STErrorMap}) \tag{12}$$

Here, M_3 measures the average size of connected components and M_4 measures the size variation of all connected components. This authentication process aims to distinguish between tampered regions caused by content-altering modifications and tampered regions caused by moderate content-preserving modifications. It is triggered when M_2 is smaller than $T_{malicious}$, which indicates that there is either a significant amount of mildly tampered error pixels (caused by moderate content-preserving modifications) or a small amount of strongly tampered error pixels (caused by minor content-altering modifications). Small M_3 and M_4 values reflect that connected components are randomly spread out, relatively small, and similar in size. As a result, we conclude that tampered regions are caused by moderate content-preserving modifications and the probe image therefore undergoes incidental attacks.

Fourth, we perform the third-level authentication using measure M_5 computed from the processed *STErrorMap* and the processed binary error map *MTErrorMap*, which is generated by marking mildly tampered error pixels in *ErrorMap* as 1's. Specifically, we remove the component(s) in *STErrorMap* and the mildly tampered error pixels in *MTErrorMap* whose distance to the centroid of the largest connected component is larger than an adaptively computed threshold *SizeC* (e.g., square root of the size of the largest connected component). The processed *STErrorMap* and *MTErrorMap* contain tampered error pixels that are near the largest connected component in *STErrorMap*. Measure M_5 re-estimates the clustering level of tampered error pixels around the largest connected component area and is computed as follows:

$$M_5 = \frac{\text{Number of error pixels in processed } \mathit{STErrorMap}}{\text{Number of error pixels in processed } \mathit{STErrorMap} \text{ and } \mathit{MTErrorMap}} \quad (13)$$

This authentication process aims to identify tampered regions when the probe image undergoes content-altering modifications together with moderate to severe content-preserving modifications. It is triggered when M_3 or M_4 is large, which indicates that a relatively large component corresponding to malicious distortion may exist. The post-processing operation removes the effect from content-preserving modifications and keeps the most important distortion area intact. Consequently, M_5 measures the clustering level of tampered error pixels in the estimated localized area. In this way, a larger M_5 indicates the presence of the maliciously tampered area.

The flowchart of computing five authentication measures is shown in Fig. 3. Extensive experiments show that the value of 0.15 for T_{median} work well on all test images and simulated attacks.

Finally, we design a quantitative method to decide the authenticity of the probe image based on the three-level five authentication measures. The flow chart of this process is summarized in Fig. 4. The involved thresholds, i.e., $T_{errorbit}$, $T_{halferrorbit}$, and $T_{malicious}$, are determined based on the predefined false negative probability of 10^{-6}. Interested readers may refer to [22] for detailed derivation of these values. In the proposed system, we set $T_{errorbit}$ as 0.4837, $T_{halferrorbit}$ as 0.2419, and $T_{malicious}$ as 0.6085.

2.5. Validation of defined error pixels and authentication measures

Fig. 5 presents the tampered error pixel distribution after performing no attack and three kinds of attacks (e.g., malicious attack, three levels of JPEG compression attacks, and malicious attack followed by each of three levels of JPEG compression attacks) on the watermarked "Lena" image, respectively. For the error pixel distribution under each attack, we sequentially display the distribution of mildly tampered error pixels (*MTErrorMap*), strongly tampered error pixels (*STErrorMap*), and their post-processed counterparts. To facilitate discussion, we also list five authentication measures for each probe image. Some of these measures are not used in the authentication process and therefore are marked as not used.

We clearly observe the following: (1) Fig. 5(a) shows that *MTErrorMap* and *STErrorMap* contain all 0's when no attack occurs to the watermarked image. In other words, watermark is successfully extracted and the probe image is authentic. (2) Fig. 5(b) shows that *MTErrorMap* and *STErrorMap* respectively contain clustered mildly and strongly tampered error pixels when the malicious attack is applied to the watermarked image. Strongly tampered pixels are also clustered within tampered areas under malicious attack. (3) Fig. 5(c), (e), and (g) show that *MTErrorMap* and *STErrorMap* respectively contain a majority of randomly spread mildly tampered error pixels and a significantly fewer number of randomly spread strongly tampered error pixels. When compression QF decreases, more mildly and strongly tampered error pixels appear in their corresponding maps. However, these error pixels tend to be isolated as both M_3 and M_4 are small. (4) Fig. 5(d), (f), and (h) show that *MTErrorMap* contains a majority of randomly spread mildly tampered error pixels resulting from compression. *STErrorMap* contains a majority of clustered strongly tampered error pixels resulting from the malicious attack together with a few randomly spread strongly tampered error pixels resulting from compression. When compression QF decreases, more mildly and strongly tampered error pixels appear in their corresponding maps. In the meantime, both M_3 and M_4 increase mainly because more strongly tampered pixels appear in the map at a higher compression ratio. However, strongly tampered pixels are clustered within tampered areas after removing outlier error pixels.

These observations verify the effectiveness of the defined tampered error pixels and authentication measures. Based on the

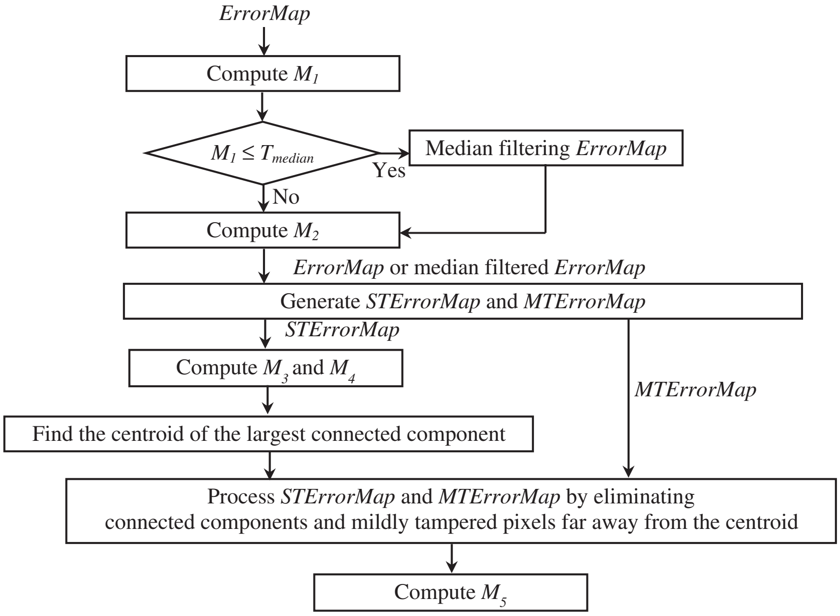

ErrorMap
Compute M_1
$M_1 \le T_{median}$
Yes
No
Median filtering ErrorMap
Compute M_2
ErrorMap or median filtered ErrorMap
Generate STErrorMap and MTErrorMap
STErrorMap
MTErrorMap
Compute M_3 and M_4
Find the centroid of the largest connected component
Process STErrorMap and MTErrorMap by eliminating connected components and mildly tampered pixels far away from the centroid
Compute M_5

Fig. 3. Flowchart of computing five authentication measures.

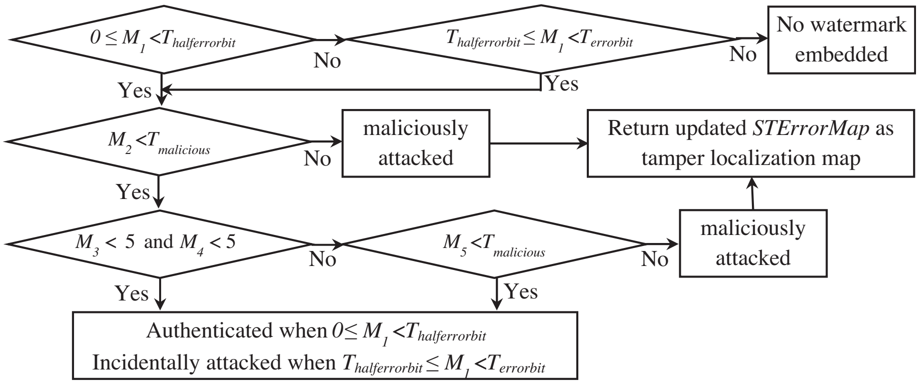
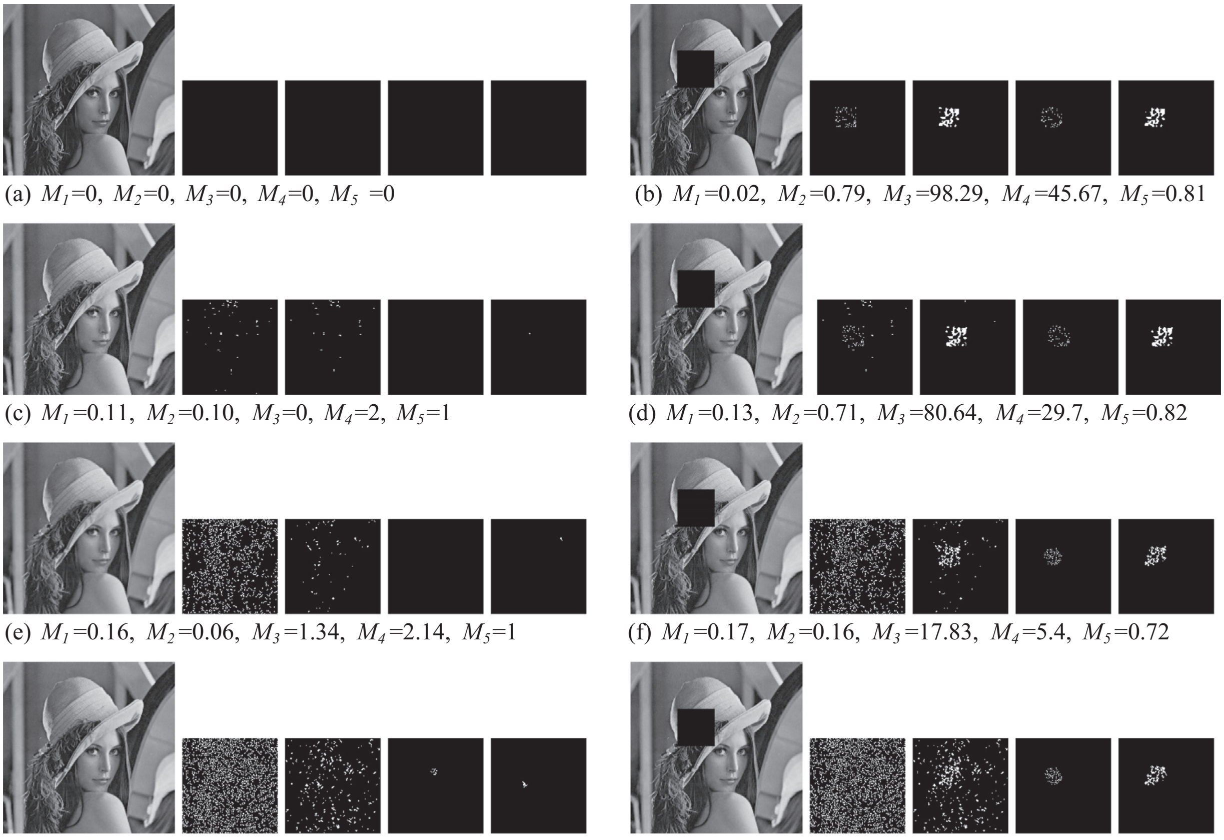

$0 \le M_1 < T_{halferrorbit}$
No
$T_{halferrorbit} \le M_1 < T_{errorbit}$
No
No watermark embedded
Yes
Yes
$M_2 < T_{malicious}$
No
maliciously attacked
Return updated *STErrorMap* as tamper localization map
Yes
$M_3 < 5$ and $M_4 < 5$
No
$M_5 < T_{malicious}$
No
maliciously attacked
Yes
Yes
Authenticated when $0 \le M_1 < T_{halferrorbit}$
Incidentally attacked when $T_{halferrorbit} \le M_1 < T_{errorbit}$

Fig. 4. Authentication decision with five authentication measures.


(a) M_1=0, M_2=0, M_3=0, M_4=0, M_5 =0

(b) M_1=0.02, M_2=0.79, M_3=98.29, M_4=45.67, M_5=0.81

(c) M_1=0.11, M_2=0.10, M_3=0, M_4=2, M_5=1

(d) M_1=0.13, M_2=0.71, M_3=80.64, M_4=29.7, M_5=0.82

(e) M_1=0.16, M_2=0.06, M_3=1.34, M_4=2.14, M_5=1

(f) M_1=0.17, M_2=0.16, M_3=17.83, M_4=5.4, M_5=0.72

(g) M_1=0.26, M_2=0.17, M_3=2.85, M_4=3.11, M_5=0.53

(h) M_1=0.27, M_2=0.21, M_3=10.19, M_4=4.07, M_5=0.71

Fig. 5. Illustration of the tampered error pixel distribution of *MTErrorMap*, *STErrorMap*, post-processed *MTErrorMap*, and post-processed *STErrorMap*, whose sizes are enlarged for easy reading. (a) Watermarked image without any attack; (b) maliciously attacked watermarked image without compression (M_3, M_4, and M_5 are not used); (c) JPEG compressed watermarked image of 80% QF (M_5 is not used); (d) maliciously attacked watermarked image with JPEG compression of 80% QF (M_3, M_4, and M_5 are not used); (e) JPEG compressed watermarked image of 70% QF (M_5 is not used); (f) maliciously attacked watermarked image with JPEG compression of 70% QF; (g) JPEG compressed watermarked image of 50% QF (M_5 is not used); (h) maliciously attacked watermarked image with JPEG compression of 50% QF.

authentication algorithm, the proposed system successfully detects watermarked image shown in Fig. 5(a) as authentic, watermarked images shown in Fig. 5(b), (d), (f), and (h) as maliciously attacked, and watermarked images shown in Fig. 5(c), (e), and (g) as incidentally attacked.

3. Performance analysis

In the following, we evaluate the performance of the proposed scheme in terms of the quality of the watermarked image, the robustness of secure watermark, and the tamper detection sensitivity.

3.1. Quality of the watermarked image

Image distortion is caused by modifications of wavelet coefficients in the embedding process. Both quantizer q and watermark payload p (e.g., equals to 1/16 of the number of pixels in an image) affect the quality of the watermarked image. A larger quantizer incurs more modification to wavelet coefficients and consequently results in more degradation in the watermarked image. Similarly, a larger payload leads to more degradation. Experimental results on 200 8-bit grayscale images show that the average *PSNR* value of their watermarked images produced by the proposed scheme is 41.39 db. This average is consistent with the expected values of 44.12, 42.66, 41.42, and 40.33 for quantizers of 11, 13, 15, and 17, respectively, and is higher than the empirical value (35.00 db) for the image without perceivable degradation [29].

3.2. Robustness of secure watermark

Since we use the same strategy to generate secure watermark in watermark embedding and authentication processes, the robustness of secure watermark is important to the proposed scheme. We clarify three aspects of the analysis process, which are the main steps of generating secure watermark, to analyze its robustness.

Using a specific non-overlapping 4 × 4 block q_Blk_i as an example, we start the explanation with the quantization of non-overlapping 8 × 8 blocks using the JPEG quantization matrix. By applying this step, modification within the range of $[-coef/2, coef/2]$ can be preserved, where *coef* is the quantization value at the same position of the quantization matrix. That is, when modifications are applied to q_Blk_i, the secure watermark bit can still be regenerated if the modification falls within the preserved range of $[-coef/2, coef/2]$. From the quantization matrix, we can tell that coefficients at different locations in q_Blk_i have a different level of robustness since their corresponding quantization values differ at different locations.

Secondly, we calculate the SV's of $subblock_{i\_1}$, $subblock_{i\_2}$, and $subblock_{i\_3}$ of block q_Blk_i. Because we generate bits B_1, B_2, and B_3 using relationships among SV's, which are more stable than the SV's themselves, it is reasonable to assume that the regenerated watermark bit will be the same as the corresponding watermark bit in the embedding process. In other words, B_1, B_2, and B_3 will not be changed even when the values of SV's can be changed because of modifications of block *Blk*.

Thirdly, we generate the watermark bit using Eq. (3) to complement possible changes in relationships encoded in B_1, B_2, and B_3. This step is to increase the chance that the regenerated watermark bit is the same as the embedded watermark bit, even when B_1, B_2, or B_3 is different from the ones generated in the embedding process. For example, changes in $S_{i\_1}(1,1)$ may lead to changes in any of two relationships, i.e., B_1 and B_3. If both relationships are changed, the regenerated watermark is the same as the embedded watermark. Even when one of the relationships (B_1 or B_3) is changed, the regenerated watermark may still stay the same if B_2 is changed. As a result, the proposed watermark generation method is robust when small incidental attacks are applied to the watermarked image. Experimental results shown in Fig. 5 and other extensive experimental results shown in Section 4 also confirm this.

Here, we use one example to illustrate the robustness of secure watermark. Fig. 6(a) shows generating the content-dependent embedded watermark bit of a 4 × 4 block, whose coefficients are the intersection of row 1–4 and column 9–12 of "Lena". So, 16 upper-left coefficients of the quantization matrix are used in the modification process. Fig. 6(b) shows regenerating the content-dependent watermark bit of the block at the same position after applying the default JPEG compression of 75% QF on the host image. We can see that even though the block has been changed after compression, the watermark bit regenerated in the authentication process is the same as the watermark bit generated in the embedding process.

3.3. Tamper detection sensitivity

The tamper detection sensitivity of the proposed scheme is determined by the quantizer q and changes in secure watermark. The error map captures changes in quantization results and makes the tampering detectable for $k \in Z$ in the following two cases:

(1) The wavelet coefficient $LL_i'(1,1)$ of the watermarked image is $2kq$, and the manipulation causes a shift of $LL_i'(1,1)$ in the range of $[(0.5+2k)q, (1.5+2k)q)$.
(2) The wavelet coefficient $LL_i'(1,1)$ of the watermarked image is $2kq+q$, and the manipulation causes a shift of $LL_i'(1,1)$ in the range of $[(1.5+2k)q, (2.5+2k)q)$.

That is, the scheme is able to detect all the changes satisfying the above two conditions. Small changes of a half of q or other changes in the range of $[(-0.5+2k)q, (0.5+2k)q]$ in the distorted area do not modify the parity of the quantized approximation value. As a result, the scheme is robust against mild to moderate content-preserving modifications. Furthermore, q is adaptive for each embedding block. Such variation reduces the possibility of misclassifying, given that the range of $[(-0.5+2k)q, (0.5+2k)q]$ where the misclassification will happen is inconsistent across the whole image.

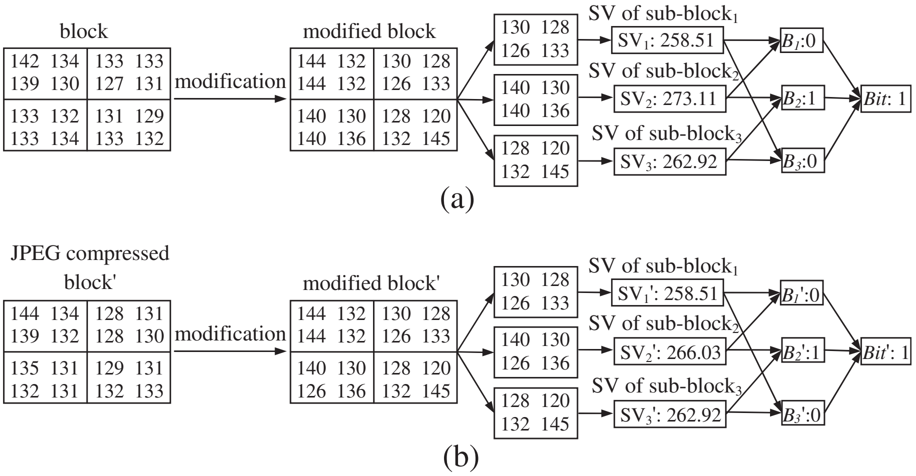


Fig. 6. Example of the robustness of secure watermark. Generate the content-dependent watermark bit of the 4 × 4 block in two processes: (a) embedding process; (b) authentication process.

The proposed scheme is also able to detect changes in three sub-blocks of each JPEG quantized block. As shown in Fig. 6, content-dependent watermark regenerated in authentication is the same as the one generated in embedding under mild incidental attacks. However, it may not be the same if moderate to severe incidental attacks or malicious attack is applied on the watermarked image. As a result, the scheme can capture relationship changes, which may correspond to content changes, among three sub-blocks.

It should be noted that some pixels in the tampered area may be missed when changes in wavelet coefficients do not satisfy the above two conditions or changes in sub-blocks cannot be captured in SVs-based relationships. The proposed authentication process may compensate this possible misclassification. Since the proposed authentication scheme uses the 3 × 3 region in *ErrorMap* to make the authentication decision, where each error pixel corresponds to 4 × 4 in the original image, the detection unit of the proposed scheme is 12 × 12.

4. Experimental results

We test the proposed scheme on five representative 8-bit 512 × 512 grayscale images and 200 8-bit grayscale images including 30 common images and 170 images converted from personally collected pictures. We perform four sets of experiments on these 200 images to compare the proposed scheme with its two variant schemes and five peer schemes using their empirically determined parameters. Variant 1 scheme embeds random instead of secure watermark in each 4 × 4 non-overlapping block. Variant 2 scheme embeds secure watermark in each 8 × 8 instead of 4 × 4 non-overlapping block. The other chosen systems include Maeno et al.'s scheme using the random bias [12], Yang and Sun's scheme [16], Che et al.'s scheme [17], Cruz et al.'s scheme [18], and Qi and Xin's scheme [22]. A more recent work [23] is not included in this comparison mainly due to its inability to process color images and the difficulty to choose three suitable parameters to make a tradeoff between the output image quality and the ability to detect tampering while preserving a specific level of robustness against image processing attacks. All peer schemes use a measure similar to the M_1's in their authentication process. Both Yang's and Qi's schemes also use another measure similar to the M_2's. In addition, both schemes summarize thresholds for detecting a probe image as authentic, incidentally distorted, or maliciously distorted. The other three schemes use the threshold of 0.3, which is a little bit higher than $T_{halferrorbit}$ (i.e., 0.2418) for M_1's in the proposed scheme. Qi's scheme is the only one that visually shows tamper localization results and lists the values of authentication measures. The other four schemes rely on visual inspection to show their effectiveness. The variant 1 system is similar to Qi's scheme except that it uses the adaptive quantizer to embed watermark and uses five measures in making authentication decision. The proposed system further embeds SV-based secure watermark instead of private-key-based random watermark to capture more distortions. In the experiments on various malicious attacks, we present both the values of authentication measures and tamper localization results to validate the effectiveness of the proposed scheme and its variant schemes.

4.1. Watermark invisibility

We compare the quality of the watermarked images produced by the eight compared schemes. Table 1 summarizes the *PSNR* values after embedding watermarks in five images and the average *PSNR* values after embedding watermarks in 200 images using each of the compared schemes. This table clearly shows that *PSNR* values of the proposed scheme and its two variants are larger than 40.00 db and are comparable with the expected *PSNR* value mentioned in Section 3.1. With the exception of Cruz's scheme [18] which embeds watermark bits in larger blocks of 16 × 16, the proposed scheme achieves a higher average *PSNR* value (listed in the last column) than the other four peer schemes.

4.2. Robustness to common image processing attacks

We perform four kinds of representative image processing attacks on 200 watermarked images to compare the robustness of the proposed scheme and five peer schemes. The results of the two variant systems are not listed here since they follow the similar trends as the proposed system. The four kinds of attacks are ten levels of image blurring using circular averaging filters, ten levels of Gaussian low-pass filtering using 3 × 3 rotationally symmetric filters, ten levels of median filtering, and five levels of salt and pepper noise. The left side of Fig. 7 shows the plot of the average M_1 values of 200 watermarked images under each image processing attack for all six schemes. The right side of Fig. 7 shows the plot of the average M_2 values of 200 watermarked images under each image processing attack for three schemes (e.g., proposed, Yang's, and Qi's schemes). We do not plot the average values of M_3, M_4, and M_5 here since none of the peer systems uses similar measures. In addition, either both M_3 and M_4 are smaller than 5 or M_5 is smaller than $T_{malicious}$ (0.6085) in all experiments. Consequently, the authentication decision is dependent on measures M_1 and M_2. To ease discussion, we focus on the comparison of these two measures.

Fig. 7 clearly shows that the proposed scheme works well under blurring, Gaussian low-pass filtering, or salt and pepper noise attacks since M_2's are below $T_{malicious}$ and M_1's are below $T_{errorbit}$. The watermarked image under median filtering with a filter size of 3–7 is detected as incidentally distorted. However, the watermarked image under median filtering with a larger filter size is detected as a non-copyrighted image. This is reasonable due to its significant changes on the watermarked image. Compared to five peer systems, the proposed scheme yields a little larger values of M_1's than Qi's scheme and yields the smallest and comparable values of M_1's among the other four schemes. It yields a little larger values of M_2's than Qi's scheme and significantly smaller values of M_2's than Yang's scheme. This indicates that the proposed scheme is comparable with Qi's scheme to capture changes caused by image processing attacks and is more robust than the other four schemes in classifying a watermarked image as authentic or incidentally distorted.

4.3. Robustness to JPEG lossy compression and JPEG2000 lossy compression attacks

We individually perform JPEG and JPEG2000 lossy compression on 200 watermarked images to compare the robustness of the

Table 1
Comparison of PSNR values.

	Lena	Peppers	Baboon	Airplane	Cameraman	Average of 200 images
Proposed	41.76	41.24	41.09	41.04	41.80	41.35
Variant 1	41.15	40.34	40.55	40.41	41.38	40.63
Variant 2	47.71	47.14	46.99	46.90	47.74	47.31
Qi	41.04	40.51	41.30	40.35	40.18	40.54
Yang	38.28	37.05	32.76	35.70	42.59	36.72
Che	39.45	37.95	37.84	37.64	38.76	37.43
Maeno	32.42	31.56	31.29	31.02	32.64	31.65
Cruz	45.78	45.29	45.09	44.51	46.10	44.32

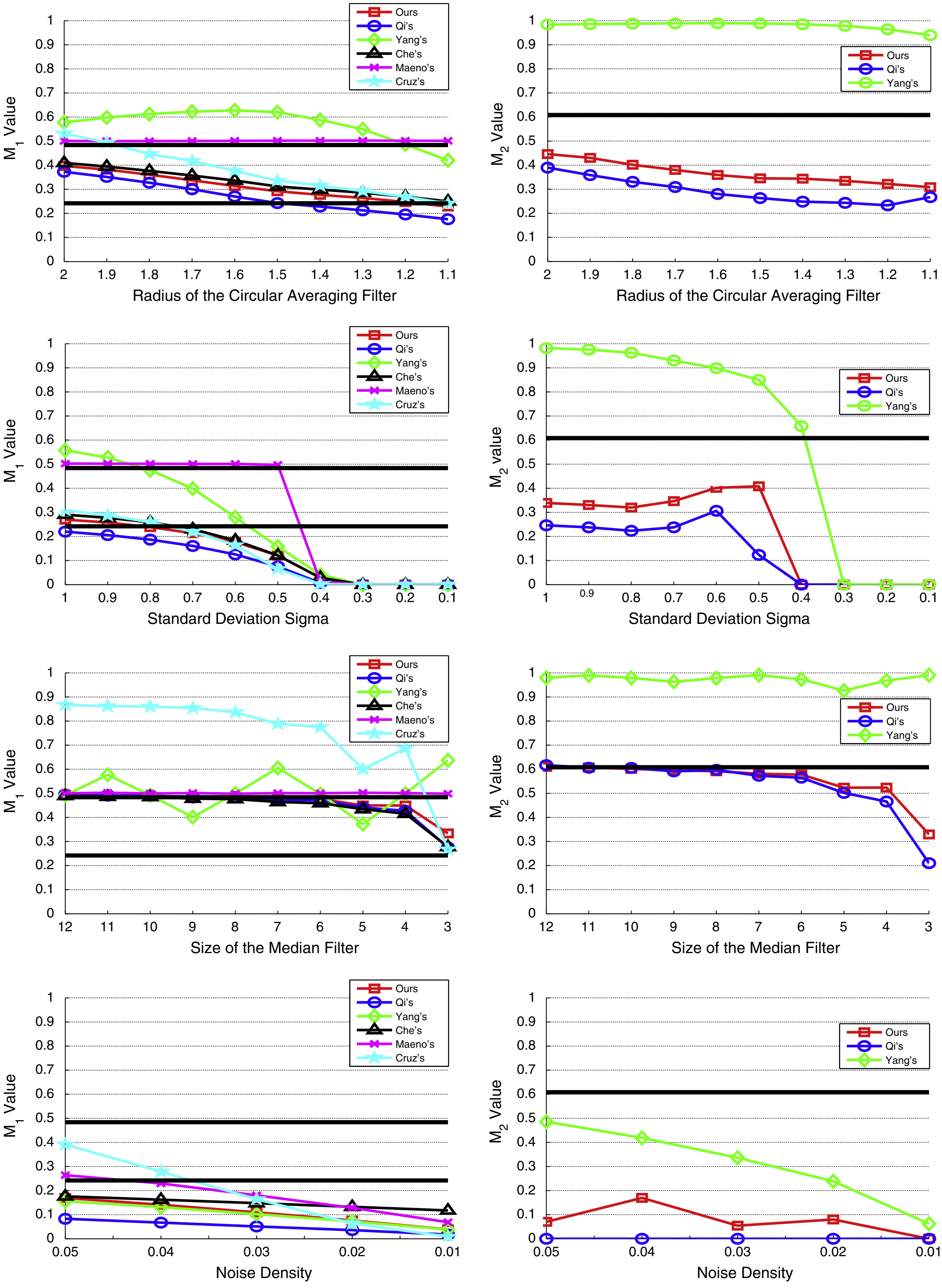


Fig. 7. Comparison of various common image processing attacks on M_1's (left) of the proposed scheme and five peer schemes and M_2's (right) of the proposed, Yang's, and Qi's schemes. 1st row: Image blurring. 2nd row: Gaussian low-pass filtering. 3rd row: Median filtering. 4th row: Salt and peppers noise.

proposed scheme and five peer schemes. The results of the two variant systems are not listed here since they follow the similar trends as the proposed system. The left plot of Fig. 8 compares the average values of M_1's of 200 watermarked images under no attack and 10 levels of JPEG compression attacks. The right plot of Fig. 8 compares the average values of M_2's of 200 watermarked images under no attack and 10 levels of JPEG compression attacks of three schemes (e.g., proposed, Qi's, and Yang's schemes). In all experiments, both M_3 and M_4 are smaller than 5. Consequently, M_5 is not used in authentication as shown in Fig. 5 and the authentication decision depends on measures M_1 and M_2.

Fig. 8 clearly shows that the proposed scheme works well under JPEG compression attacks since M_2's are below 0.6085 for a QF down to 10%. However, it detects the watermarked image under a QF of 10% as a non-copyrighted image. This is reasonable due to its significant changes on the watermarked image when large compression occurs. Compared to five peer systems, the proposed scheme yields a little larger values of M_1's than Qi's scheme and yields the smallest values of M_1's for JPEG QFs down to 30% among the other four schemes. It yields a little larger values of M_2's than Qi's scheme and significantly smaller values of M_2's than Yang's scheme. This indicates that the proposed scheme is comparable with Qi's scheme and is more robust than the other four schemes in classifying a watermarked image as authentic or incidentally distorted under JPEG compression attacks. The other four peer systems detect the watermarked images under 10–50% JPEG QFs as maliciously distorted.

To evaluate the robustness of the proposed scheme to JPEG2000 lossy compression attacks, we compare the average values of M_1's and M_2's of 200 watermarked images under 10 levels of JPEG2000 compression and their equivalent JPEG compression as shown in Fig. 9. We also plot the values of M_1's and M_2's under each attack after adding or subtracting the standard deviation values (STDV) from their average values. Since the authentication decision depends on measures M_1 and M_2, we do not plot the other three measures.

Fig. 9 clearly shows M_1's and M_2's for JPEG2000 compression attacks are much smaller than the ones for JPEG compression attacks. The relationship holds true for the average values of M_1's and M_2's adding or subtracting their corresponding STDVs. In addition, M_2's are below 0.6085 and M_1's are below 0.2418 for all JPEG2000 compressions except the ones with the QF of 100%, 200%, and 300%. The experimental results demonstrate that the proposed scheme is more robust against JPEG2000 compression than JPEG compression since it works in the same wavelet domain as the JPEG2000 compression.

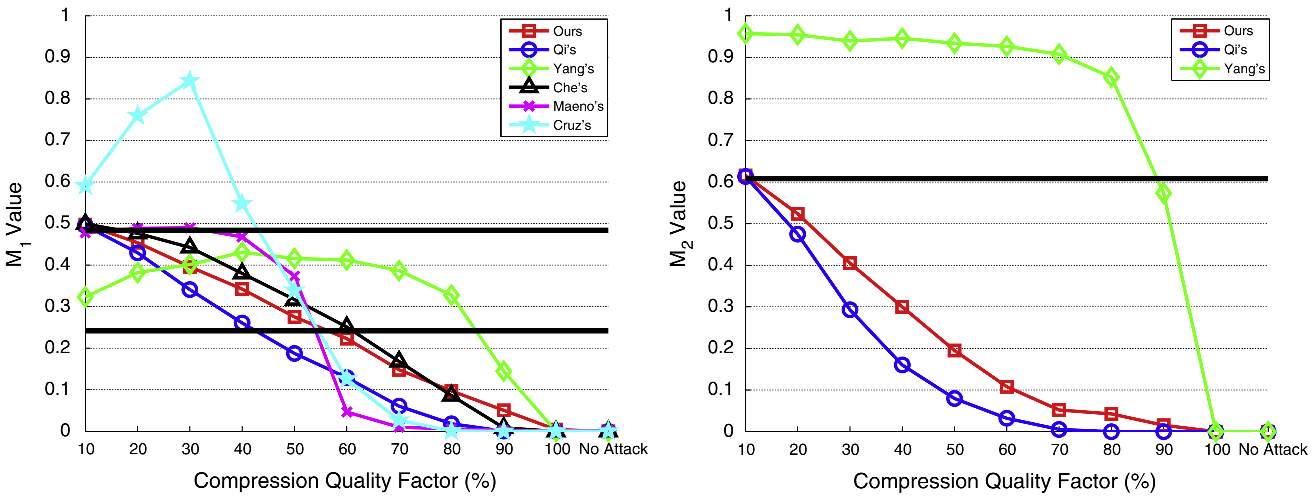


Fig. 8. Comparison of various JPEG compression attacks on M_1's (left) of the proposed scheme and five peer schemes and M_2's (right) of proposed, Qi's, and Yang's schemes.

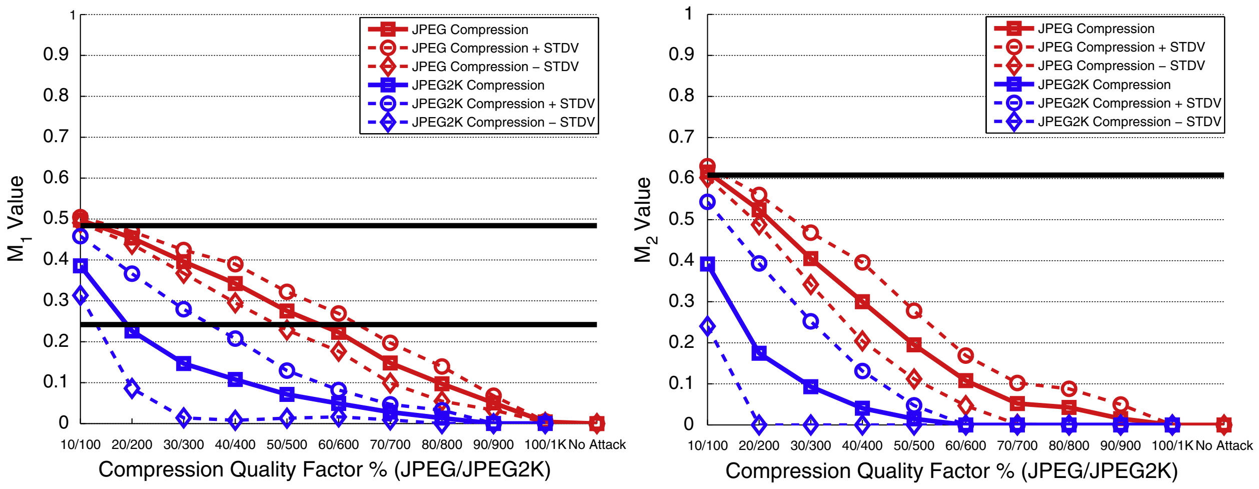


Fig. 9. Comparison of JPEG 2000 compression and their corresponding JPEG compression attacks on M_1's (left) and M_2's (right) of the proposed scheme.

4.4. Fragility to malicious attacks

We perform malicious attacks on 200 watermarked images to compare the fragileness of compared systems and demonstrate the effectiveness of the proposed scheme in localizing tampered regions.

In the first experiment, we add an irregular shape (3437 pixels) of three kinds of gray-level intensities (i.e., black, gray, and white) to the watermarked "Lena" image, wherein white is the most dissimilar to the background intensity and gray is the most similar to the background intensity. We deliberately do not apply compression to separate out its effect. Fig. 10 shows tamper localization results in yellow and lists applicable M_1 and M_2 values in a pair for each of the eight compared schemes to facilitate comparison. The other three measures (M_3, M_4, and M_5) are not listed here since the decision can be made at the first authentication level using measures M_1 and M_2. Fig. 10(a) shows the results of the proposed scheme, its two variant schemes, and Qi's scheme, which achieve the best robustness against image processing and compression attacks. Fig. 10(b) shows the results of the other four peer systems.

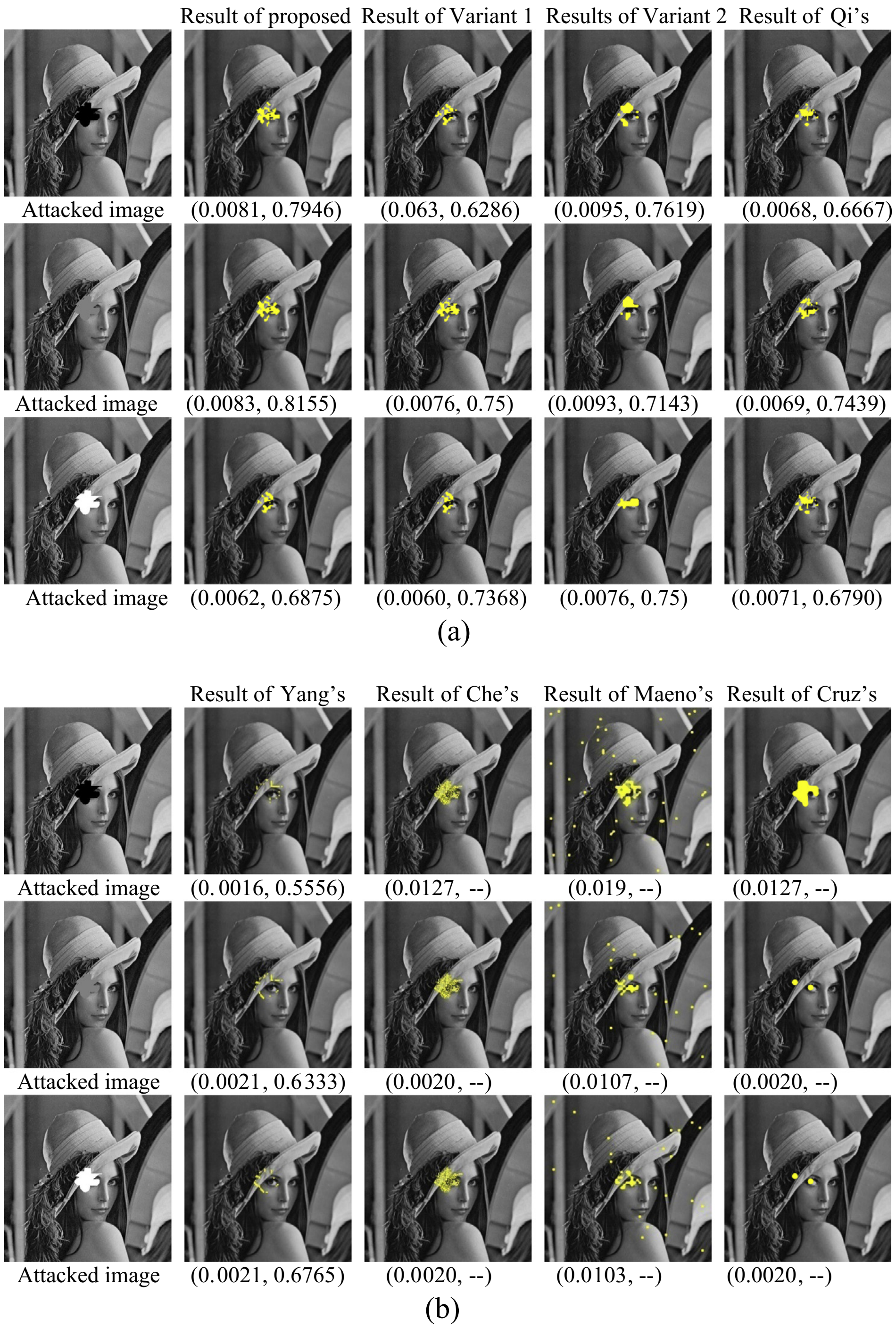


Fig. 10. Comparison of tamper localization results of eight semi-fragile watermarking schemes after adding an irregular shape of three kinds of gray-level intensities without compression. (a) Comparison of the proposed scheme, its two variants, and Qi's scheme. (b) Comparison of the other four peer systems.

Fig. 10 clearly shows that the proposed scheme and its variant schemes achieve similar localization results as Qi's and Che's schemes and outperform the other three schemes by correctly localizing tampered regions regardless of the intensity of the added shape. Based on the authentication algorithm, we conclude that the proposed, two variant, and Qi's schemes detect three attacked images as maliciously tampered and correctly localize their tampered regions. On average, the proposed scheme and its variant 2 capture more distorted pixels and yields higher value for M_2's than Qi's scheme and its variant 1. As a result, they achieve better and more reliable localization results partly due to the embedding of SV-based secure watermark, which is more fragile to content altering operations. Yang's scheme does not produce decent localization results under any malicious attack. The other three peer schemes obtain small values for M_1's, which are similar to the values obtained under image processing and compression attacks. As a result, they detect these attacked "Lena" images as incidentally distorted based on the equivalent predefined thresholds.

In the second experiment, we add the same irregular shape of a much smaller size (384 pixels) in white or gray color at different locations (e.g., smooth region in Pepper, Baboon, and Cameraman images and complex texture region in Lena and Airplane images) of five representative watermarked images. Similar to the setting in the first experiment, we deliberately do not apply compression attacks. For the same watermarked image, we add the small irregular shape at the same location so tampered regions can be easily compared. Fig. 11 shows maliciously attacked images (2nd and 5th columns) together with their corresponding tamper localization results (3rd and 4th columns, and 6th and 7th columns) produced by the proposed scheme and its variant 1, which achieve the best results in the first experiment. The results from its variant 2 are not good since embedding takes place in a large block of size 8 × 8 and the modified region is small. Fig. 11 also lists the values of authentication measures used in decision making below each localization result. It clearly shows that the proposed scheme successfully localizes small tampered regions and detects these maliciously attacked watermarked images as tampered regardless the modification locations and intensities. However, its variant scheme 1 fails to detect maliciously attacked "Pepper" image as tampered partly due to the embedding of random watermark instead of SV-based secure watermark.

In the third experiment, we apply three kinds of realistic modifications on the watermarked "Lena" image by using Photoshop to insert a decorative flower on the hat, modify the right eye, and remove the white and gray wavy decoration at the lower right corner, respectively. The modified "Lena" image is then saved as a JPEG image using the default compression setting. Fig. 12 shows tamper localization results in yellow and lists applicable M_1 and M_2 values to facilitate comparison of six schemes. The figure clearly shows that the proposed scheme achieves the best and the cleanest localization results, Qi's scheme ranks the second, and Maeno's scheme ranks the third with a few small isolated distorted regions resulting from compression. Che's and Maeno's schemes achieve comparable localization results except that it detects more distorted regions due to its less robustness to compression. Based on the authentication algorithm, we conclude that

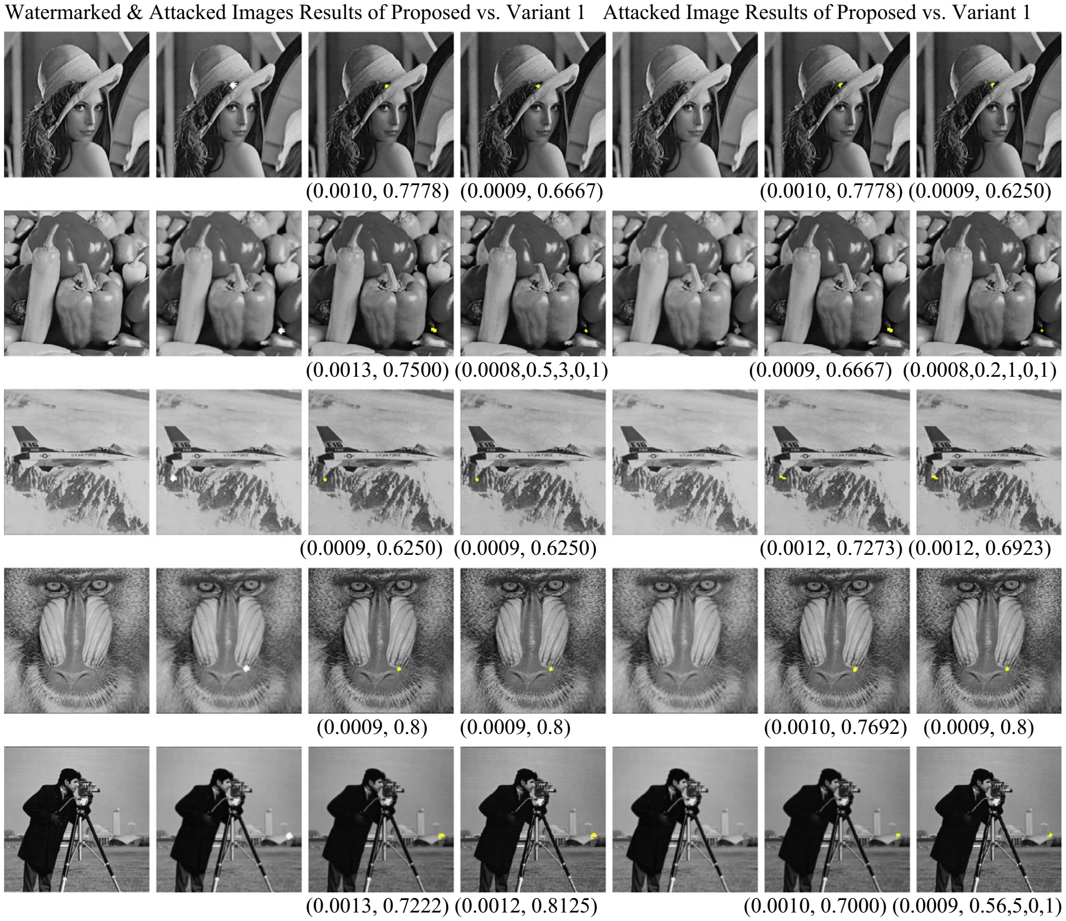


Fig. 11. Illustration of tamper localization results produced by the proposed scheme and its variant scheme 1 after adding a small irregular shape of two kinds of gray-level intensities without compression.

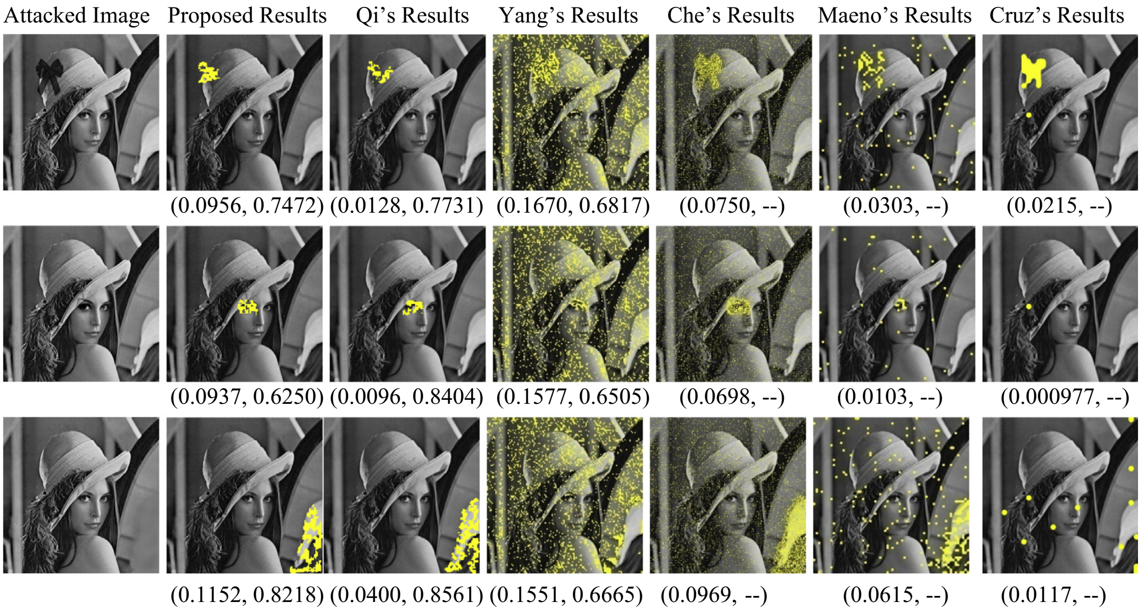


Fig. 12. Comparison of tamper localization results of six semi-fragile watermarking schemes under three realistic malicious attacks (modified by Photoshop).

both proposed and Qi's schemes detect all three modified watermarked "Lena" images as maliciously tampered and correctly localize tampered regions. On average, the proposed scheme captures more distorted pixels than Qi's scheme. As a result, it achieves better and more reliable localization results than Qi's scheme. Yang's scheme detects these modified watermarked images as maliciously distorted. However, it does not produce decent localization results due to less robustness to compression. The other three schemes obtain small values for M_1's, which are similar to the values obtained under common image processing

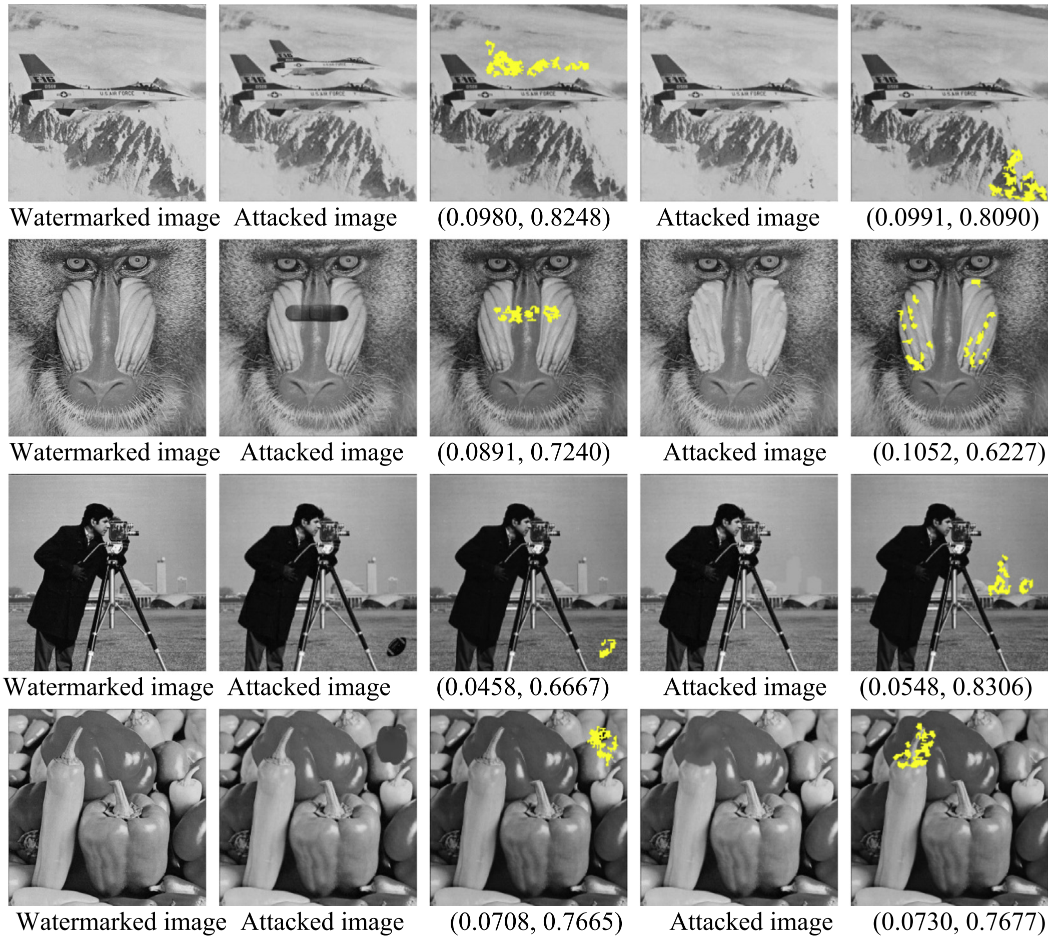


Fig. 13. Illustration of the tamper localization results of the proposed scheme under realistic malicious attacks (modified by Photoshop).

attacks. As a result, they detect these maliciously attacked images as incidentally distorted based on the equivalent predefined thresholds.

In the fourth experiment, we use Photoshop to insert an external object and remove an object at different locations of four additional watermarked images. Fig. 13 shows maliciously attacked images (2nd and 4th columns) together with their corresponding tamper localization results (3rd and 5th columns) produced by the proposed scheme. It also lists the values of authentication measures used in decision making below each localization result. It clearly shows that the proposed scheme successfully localizes tampered regions and detects all these modified watermarked images as maliciously tampered.

In the last experiment, we use Matlab to insert a decorative flower on the hat of the watermarked "Lena" image and then save the distorted image using different QFs ranging from 80% down to 30% with a step size of 5%. We compare tamper detection results of the proposed scheme with two schemes (e.g., variant scheme 1 and Qi's scheme), which achieve better localization results than other compared systems. Fig. 14 shows the sample maliciously attacked image (1st column) and tamper localization results produced by the proposed scheme and its variant scheme 1 for 11 JPEG compressed maliciously attacked images and tamper localization results produced by Qi's scheme for 5 JPEG compressed maliciously attacked images. It clearly shows that the proposed scheme, the variant scheme 1, and Qi's scheme successfully localize tampered regions for a QF down to 30%, 35%, and 65%, respectively. The proposed scheme also identifies significantly more distorted pixels than the other two compared schemes under the same compression QF.

To quantitatively compare the performance of three schemes, we also list the values of their applicable authentication measures in Table 2. It shows more error pixels are detected by three schemes for lower compression QF. The proposed scheme identifies a significantly more error pixels, which correspond to a larger value of M_1, than the other two schemes for the same QF. This is mainly due to its ability to capture changes both in wavelet coefficients and in SV-based relationships. The proposed scheme also tends to produce a smaller value of M_2. However, after applying post-processing, it produces large values for M_3, M_4, and M_5 for all maliciously attacked images. Based on the authentication algorithm, we conclude that the proposed scheme, the variant 1 scheme, and Qi's scheme detect modified watermarked images with JPEG compression of a QF down to 30%, 55%, and 65% as maliciously tampered, respectively, which are in accordance with the visual results shown in Fig. 14. This is a significant improvement since its variant 1 scheme, Qi's scheme, and the other four peer schemes may fail to detect tamper areas for a QF of less than 55%, less than 65%, and around 75% (shown in Fig. 10), respectively. In other words, the proposed scheme is more robust than the other compared systems in detecting tamper localization results under various compression QFs. However, the proposed scheme may lead to wrong authentication results when a high compression ratio is used to compress distorted images. For example, tamper localization results for JPEG compression QF of 25% are incorrect.

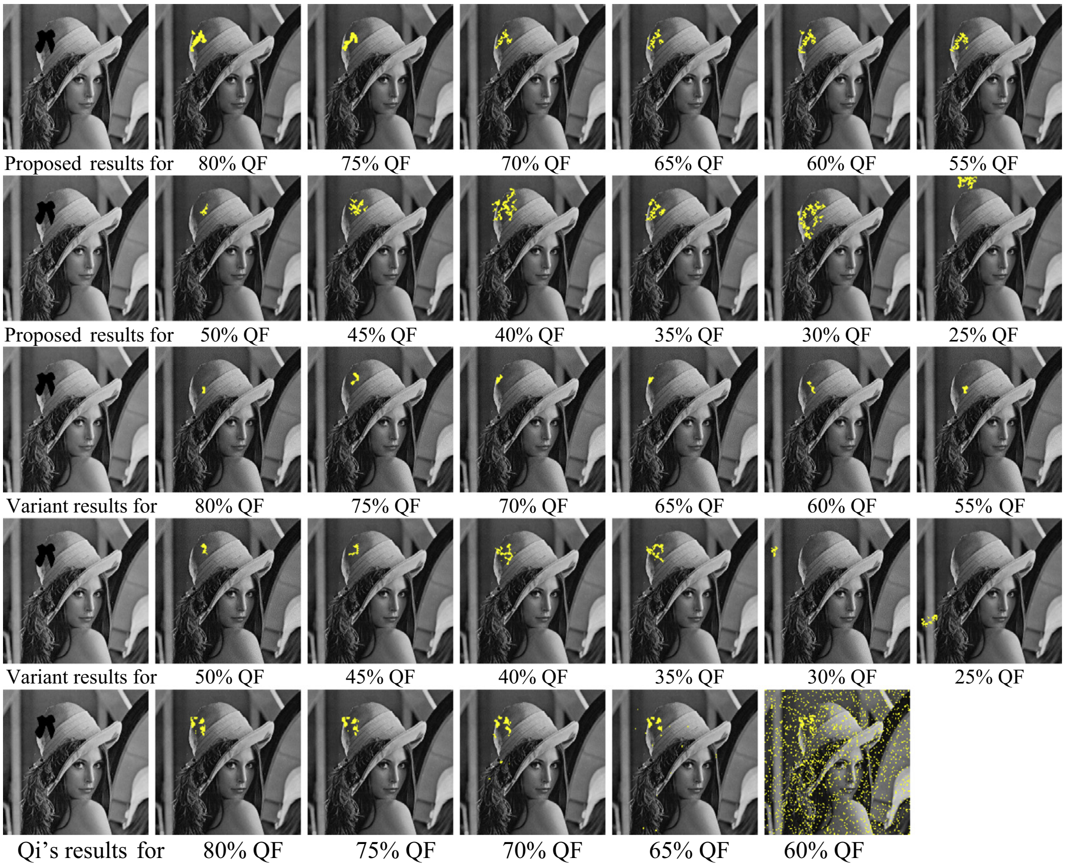


Fig. 14. Comparison of tamper localization results of the proposed scheme, its variant 1 scheme, and Qi's scheme under a realistic malicious attack followed by JPEG compression of different QFs.

Table 2
Comparison of authentication measures of the proposed scheme, its variant 1 scheme, and Qi's scheme under malicious attacks with different levels of JPEG compression.

QF (%)	M_1			M_2			M_3		M_4		M_5	
	Prop.	Var.	Qi's	Prop.	Var.	Qi's	Prop.	Var.	Prop.	Var.	Prop.	Var.
80	0.12	0.02	0.02	0.58	0.75	0.75	17.83	14.00	35.92	2.16	0.9	1
75	0.14	0.04	0.04	0.52	0.76	0.74	32.42	21.67	42.15	8.33	0.90	0.97
70	0.16	0.06	0.06	0.09	0.66	0.67	3.31	9.71	8.12	6.85	0.66	0.83
65	0.19	0.09	0.09	0.10	0.59	0.66	2.90	9.17	6.16	6.77	0.70	0.79
60	0.21	0.12	0.11	0.13	0.52	0.06	3.45	17.00	7.79	7.87	0.63	0.83
55	0.25	0.14	–	0.16	0.53	–	3.23	13.17	6.13	7.03	0.63	1
50	0.27	0.16	–	0.18	0.07	–	3.40	3.00	5.51	4.38	0.62	0.68
45	0.29	0.19	–	0.19	0.08	–	3.42	2.90	5.01	4.20	0.66	0.60
40	0.32	0.24	–	0.26	0.13	–	3.89	3.10	7.40	6.35	0.64	0.58
35	0.35	0.28	–	0.32	0.18	–	4.77	3.17	6.04	5.48	0.63	0.55
30	0.39	0.32	–	0.38	0.25	–	5.91	3.84	10.16	3.74	0.69	0.61
25	0.42	0.37	–	0.45	0.35	–	7.81	5.42	10.41	6.18	0.63	0.55

5. Conclusions

We present an effective SV-based semi-fragile watermarking scheme for image content authentication with tamper localization. The contributions of the proposed scheme are:

- Utilizing relationships of SVs among three sub-blocks of each 4×4 non-overlapping JPEG quantized block to extract content-dependent watermark in both watermark embedding and authentication processes. This extracted watermark is able to capture the changes in SV-based relationships that likely correspond to content changes in non-embedding sub-blocks.
- Generating secure watermark by performing the logical "*xor*" operation on SV-based content-dependent and private-key-based content-independent watermark. This secure watermark encodes image content and therefore is more fragile to content-altering operations and moderate to severe content-preserving operations.
- Merging relationships of SVs among three sub-blocks of each 4×4 non-overlapping JPEG quantized block to choose its adaptive quantizer q for both embedding and extraction processes. Such variation reduces the possibility of misclassification and improves the quality of the watermarked image.
- Applying the adaptive quantization method to embed secure watermark in the wavelet domain so that a majority of image distortions, which cause the intensity shift by a value larger than a half of q or cause image content to change, can be detected in the authentication process.
- Defining a three-level authentication process involving five measures to quantitatively detect the authenticity of the probe image and prove tampering, with M_1 measuring the similarity between extracted and embedded watermarks, M_2 measuring the clustering level of tampered error pixels, M_3 measuring the average size of connected components formed by strongly tampered error pixels, M_4 measuring the variation in the sizes of connected components, and M_5 measuring the clustering level of tampered error pixels around the potential distorted area.
- Using five authentication measures derived from three binary maps (e.g., *ErrorMap*, *STErrorMap*, and *MTErrorMap*) to compensate possible misclassification and ensure that the proposed scheme is capable of capturing distortions and localizing tampered areas when a moderate noise signal or a higher compression is also involved.

Extensive experimental results show that the proposed scheme successfully distinguishes malicious attacks from non-malicious tampering of image content. It also accurately localizes maliciously tampered regions even under moderate to severe JPEG compressions. The proposed scheme is more robust to various acceptable content-preserving operations and more fragile to malicious distortions than five semi-fragile watermarking schemes and its two variant schemes. It can be easily extended to color images.

Our future work includes studying different authentication measures, addressing geometric attack issues, and testing more images of various types.

References

[1] C. Rey, J.L. Dugelay, A survey of watermarking algorithms for image authentication, EURASIP J. Appl. Signal Process. 6 (2002) 613–621.

[2] X. Zhang, S. Wang, Z. Qian, G. Feng, Reversible fragile watermarking for locating tampered blocks in JPEG images, Signal Process. 90 (2010) 3026–3036.

[3] X. Zhang, S. Wang, Z. Qian, G. Feng, Reference sharing mechanism for watermark self-embedding, IEEE Trans. Image Process. 20 (2) (2011) 485–495.

[4] C. Qin, C. Chang, P. Chen, Self-embedding fragile watermarking with restoration capability based on adaptive bit allocation mechanism, Signal Process. 92 (2012) 1137–1150.

[5] C. Qin, C. Chang, K. Chen, Adaptive self-recovery for tampered images based on VQ indexing and inpainting, Signal Process. 93 (2013) 933–946.

[6] M. Tsai, C. Chien, Authentication and recovery for wavelet-based semifragile watermarking, Opt. Eng. 47 (6) (2008) 1–10. 067005.

[7] R. Chamlawi, A. Khan, Digital image authentication and recovery: employing integer transform based information embedding and extraction, Inf. Sci. 180 (2010) 4909–4928.

[8] C. Li, B. Ma, Y. Wang, D. Huang, Z. Zhang, A secure semi-fragile self-recoverable watermarking algorithm using group-based wavelet quantization, in: Advances in Multimedia Information Processing, Lecture Notes in Computer Science, vol. 7674, 2012, pp. 327–336.

[9] Y. Huo, H. He, F. Chen, A restorable semi-fragile watermarking combined DCT with interpolation, in: Digital Forensics and Watermarking, Lecture Notes in Computer Science, vol. 8389, 2014, pp. 393–408.

[10] I. Cox, M. Miller, J. Bloom, J. Fridrich, T. Kalker, Digital Watermarking and Steganography, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2008.

[11] C. Ho, C. Li, Semi-fragile watermarking scheme for authentication of JPEG images, in: Proceedings of Int. Conf. on ITCC, 2004, pp. 7–11.

[12] K. Maeno, Q. Sun, S. Chang, M. Suto, New semi-fragile image authentication watermarking techniques using random bias and nonuniform quantization, IEEE Trans. Multimedia 8 (1) (2006) 32–45.

[13] X. Zhou, X. Duan, D. Wang, A semi-fragile watermarking scheme for image authentication, in: Proceedings of the 10th Int. Conf. on Multimedia Modeling, 2004, pp. 374–377.

[14] Y. Hu, D. Han, Using two semi-fragile watermarks for image authentication, in: Proceedings of the 4th Int. Conf. on Machine Learning and Cybernetics, 2005, pp. 5484–5489.

[15] Y. Zhu, C. Li, H. Zhao, Structural digital signature and semi-fragile fingerprinting for image authentication in wavelet domain, in: Proceedings of IAS, 2007, pp. 478–483.

[16] H. Yang, X. Sun, Semi-fragile watermarking for image authentication and tamper detection using HVS model, in: Proceedings of Int. Conf. on Multimedia and Ubiquitous Engineering, 2007, pp. 1112–1117.

[17] S. Che, B. Ma, Z. Che, Semi-fragile image watermarking algorithm based on visual features, in: Proceedings of Int. Conf. on Wavelet Analysis and Pattern Recognition, 2007, pp. 382–387.

[18] C. Cruz, R. Reyes, M. Nakano, H. Perez, Image content authentication system based on semi-fragile watermarking, in: Proceedings of the 51st Midwest Symposium on Circuits and Systems, 2008, pp. 306–309.

[19] X. Zhang, L. Cui, L. Shao, A fast semi-fragile watermarking scheme based on quantizing the weighted mean of integer Haar wavelet coefficients, in: Proceedings of Symposium on Photonics and Optoelectronics, 2012, pp. 1–4.

[20] R.O. Preda, Semi-fragile watermarking for image authentication with sensitive tamper localization in the wavelet domain, Measurement 46 (2013) 367–373.

[21] Y. Huo, H. He, F. Chen, A semi-fragile image watermarking algorithm with two-stage detection, Multimedia Tools Appl. 72 (2014) 123–149.

[22] X. Qi, X. Xin, A quantization-based semi-fragile watermarking scheme for image content authentication, J. Vis. Commun. Image Represent. 22 (2011) 187–200.

[23] H.M. Al-Otum, Semi-fragile watermarking for grayscale image authentication and tamper detection based on an adjusted expanded-bit multiscale quantization-based technique, J. Vis. Commun. Image Represent. 25 (2014) 1064–1081.

[24] B. Chen, G.W. Wornell, Quantization index modulation: a class of provably good methods for digital watermarking and information embedding, IEEE Trans. Inform. Theory 47 (2001) 1423–1443.

[25] T. Liu, Z. Qiu, The survey of digital watermarking-based image authentication techniques, in: Proceedings of the 6th Int. Conf. on Signal Processing, 2002, pp. 1556–1559.

[26] M. Matsumoto, T. Nishimura, Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator, ACM Trans. Model. Comput. Simul. 8 (1) (1998) 3–30.

[27] Z.Q. Hong, Algebraic feature extraction of image for recognition, Pattern Recogn. 24 (3) (1991) 211–219.

[28] X. Qi, J. Qi, A robust content-based digital image watermarking scheme, Signal Process. 87 (6) (2007) 1264–1280.

[29] M. Hsieh, D. Tseng, Perceptual digital watermarking for image authentication in electronic commerce, Electron. Commer. Res. 4 (2004) 157–170.