Mean-quantization-based fragile watermarking for image authentication

Gwo-Jong Yu
National Central University
Department of Computer Science and
Information Engineering
Chung-Li, Taiwan

Chun-Shien Lu
Hong-Yuan Mark Liao
Academia Sinica
Institute of Information Science
Taipei, Taiwan
E-mail: liao@iis.sinica.edu.tw

Abstract. The authors propose an image authentication scheme, which is able to detect malicious tampering while tolerating some incidental distortions. By modeling the magnitude changes caused by incidental distortion and malicious tampering as Gaussian distributions with small and large variances, respectively, they propose to embed a watermark by using a mean-quantization technique in the wavelet domain. The proposed scheme is superior to the conventional quantization-based approaches in credibility of authentication. Statistical analysis is conducted to show that the probabilities of watermark errors caused by malicious tampering and incidental distortion will be, respectively, maximized and minimized when the new scheme is applied. Experimental results demonstrate that the credibility of the method is superior to that of the conventional quantization-based methods under malicious attack followed by an incidental modification, such as JPEG compression, sharpening or blurring.
[DOI: 10.1117/1.1384885]

Subject terms: fragile watermarking; image authentication; fragility; robustness; wavelet transform; human visual system.

Paper 200367 received Sep. 18, 2000; revised manuscript received Feb. 5, 2001; accepted for publication Feb. 9, 2001.

1 Introduction

The invention of the Internet provides a brilliant way of transmitting digital media. When digital media contain important information, their credibility must be ensured. As a consequence, a reliable media authentication system is indispensable when digital media are transmitted over a network. In order to save bandwidth and storage space, digital media are usually transmitted or stored in a compressed format. In addition, media such as images may be processed by blurring or sharpening for specific purposes. Under these circumstances, an image authentication system should be able to tolerate some commonly used incidental modifications, such as JPEG compression, sharpening, and/or blurring. In this paper, we focus our discussion on image authentication.

In the literature, image authentication methods can be roughly classified as being either digital-signature-based or watermark-based. The digital-signature approach[1–4] does not modify the content of an image. Instead, it extracts either global features or relational features from media for authentication purposes. For example, Bhattacharjee and Kutter[1] used the positions of a set of feature points as a digital signature. By examining the existence of feature points, images can be authenticated. Lin and Chang[4] computed the invariant relations between the coefficients of two randomly selected DCT blocks and then used them as a digital signature. Their method is able to resist JPEG compression with compression ratios (CRs) up to 20 : 1. The major limitation of a digital-signature-based method is that it can only be used for the purpose of verification, not copyright protection.

In the watermark-based image authentication approaches, on the other hand, detection of tampering is based on the fragility of a hidden watermark.[5–10] In Kundur and Hatzinakos's[5] quantization-based method, a watermark value is encoded by modulating a selected wavelet coefficient into a quantized interval. Basically, the quantity they used for modulation, which is monotonically increased from high resolution to low resolution, violates the capacity constraint of the human visual system.[11] They defined a tamper assessment function (TAF), which is the ratio of the number of tampered coefficients to the total number of coefficients in a specific subband, in order to measure the degree of tampering. They also point out if the TAF values decrease monotonically from high resolution to low resolution, then it is very likely that the manipulation is JPEG compression. However, they did not address the situation in which an instance of malicious tampering and an incidental manipulation are imposed simultaneously.

In Ref. 2, Dittmann et al. mentioned that incidental distortions, such as JPEG compression, blurring, or sharpening, should not be treated as malicious tampering. They also mentioned that if a watermarked image is tampered with maliciously, then the portions where the watermark errors emerge should be the manipulated areas. Their argument is only partially true, because incidental operations that are not malicious also cause watermark errors. Under these circumstances, one cannot judge whether a modification is malicious or not simply by looking at watermark errors. The objective of this paper is to increase the credibility of the embedded watermark by maximizing and minimizing, respectively, the probabilities of watermark errors caused by malicious tampering and incidental distortion. Under these circumstances, an instance of malicious tampering can be easily distinguished from an incidental modification. In general, the probability of watermark error caused by an incidental distortion can be reduced by either enlarging the quantization interval or reducing the quantity of modifications on coefficients. However, it is well known that the maximum quantization interval should be bounded by the human visual system[11] so that visual quality can be maintained. As a consequence, the only methodology that we can adopt here is to increase the robustness by decreasing the variance of coefficients. Owing to the fact that the variance of a subblock mean is smaller than that of an individual sample, we know that a watermark value encoded by quantizing the mean of a set of coefficients is more robust than one encoded by quantizing a single coefficient.

Under a reasonable assumption that the number of modifications caused by an incidental distortion is smaller than that caused by a malicious distortion, the modifications caused by an incidental distortion or an instance of malicious tampering can be, respectively, modeled as a Gaussian distribution with smaller or larger variance. In a good image authentication system, it is expected that the embedded watermark should be robust enough to tolerate incidental distortion and fragile enough to detect malicious tampering. However, it is also well known that robustness and fragility are two factors that compete against each other. Therefore, we need to seek a trade-off between them that can lead to the best outcome. In order to achieve that goal, a mechanism that can be used to encode a watermark so that the probabilities of watermark errors caused by malicious tampering and incidental distortion are, respectively, maximized and minimized is indispensable.

In this paper, we propose a mean-quantization-based fragile-watermarking approach that can be used to judge the credibility of a suspect image. The approach embeds a watermark by taking the mean value of a set of wavelet coefficients. Through theoretical analysis of the probabilities of watermark errors caused by malicious tampering and incidental distortion, the best number of coefficients needed to embed a watermark at each scale can be computed so that the trade-off between robustness and fragility can be optimized. Since the probability of watermark errors caused by incidental distortion at each scale is different, the detection responses at all scales should be integrated so as to obtain a global estimation of the maliciously attacked area. Then, we can use some decision rules to judge whether a suspect image has been tampered with or not.

The remainder of this paper is organized as follows. The mean-quantization-based fragile-watermarking approach is described in Sec. 2. An information-fusion technique that can be used to integrate the detection results at multiple scales is addressed in Sec. 3. Experimental results and conclusions are given in Sec. 4 and Sec. 5, respectively.

2 Mean Quantization: A New Mechanism to Achieve Better Authentication

In this section, we describe the proposed mean-quantization-based fragile-watermarking approach. In order to protect the original source, our watermark extraction process is designed in a blind-detection manner. Blind detection means the original source is not required for watermark extraction. Among the existing blind watermarking schemes, the quantization-based watermarking approach is the simplest one that achieves the goal. This is because in a quantization-based approach, a watermark is encoded and decoded by the same quantization operation. In the following, we first introduce the conventional quantization-based approach and point out its disadvantages. Then, a mean-quantization-based approach is proposed to eliminate these disadvantages. Finally, we propose a systematic way to determine an optimal number of coefficients for mean quantization.

2.1 *Disadvantages of the Conventional Quantization-Based Scheme*

The quantization-based fragile-watermarking approach[5] divides a real-number axis in the wavelet domain into intervals with equal size at each scale and assigns watermark symbols to each interval periodically. Assuming that x is a wavelet coefficient, and that q is the size of a quantization interval, the watermark symbol, which is either 0 or 1, is determined by a quantization function Q, where

$$Q(x,q)=\begin{cases}0 & \text{if } tq\leqslant x<(t+1)q \quad \text{for } t=0,\pm 2,\pm 4,\ldots,\\ 1 & \text{if } tq\leqslant x<(t+1)q \quad \text{for } t=\pm 1,\pm 3,\pm 5,\ldots.\end{cases} \tag{1}$$

Let w denote the target watermark value that is to be encoded for a wavelet coefficient x. The encoding rule is as follows: If $Q(x,q)=w$, then no modification is necessary for x; otherwise, x is updated to x^* by

$$x^*=\begin{cases}x+q & \text{if } x\leqslant 0,\\ x-q & \text{if } x>0.\end{cases} \tag{2}$$

Kundur and Hatzinakos's approach[5] uses $\delta 2^l$ as the size of a quantization interval, where δ is a prespecified positive integer, $l=1,\ldots,L$, and L is the number of scales used in the wavelet transform. In their approach, the size of a quantization interval increases monotonically from high frequency to low frequency. However, this kind of design violates the characteristics of the human visual system.[11] In our design, we take the limitations of the human visual system into consideration. On the other hand, since any modification applied to an image will change its wavelet coefficients, it is reasonable to expect that their corresponding watermark symbols will be changed, too. By comparing the extracted watermark values with the original hidden ones, the maliciously attacked area can be located. Although the fragility of the watermark proposed in Ref. 5 is able to reveal malicious tampering, that watermark is not robust enough to tolerate incidental distortions. Therefore, we seriously address this problem as well.

2.2 *The Proposed Scheme*

Watson et al.[11] investigated the sensitivity of the human eye and then proposed a wavelet-based human visual system (HVS). According to the HVS, the wavelet coefficients can be modified without causing visual artifacts. In order for a watermarked image to satisfy the transparency requirement, the quantization interval will be defined as the maximally allowable modification quantity based on the

HVS of Ref. 11. Our basic concept is that if the modification quantity of a wavelet coefficient does not exceed its corresponding masking threshold, then this modification will not raise visual awareness. Otherwise, we can say the modification is a malicious one.

Statistically, the mean value of a set of samples has variance smaller than that of a single sample. We expect that if the watermark is embedded by modulating the mean value rather than a single coefficient, the probability of watermark errors will be smaller. This is because the mean value is more difficult to move beyond the quantization interval where it is originally located. For a specific subband, let the size of a quantization interval be denoted as q, and let a set of n wavelet coefficients be denoted as x_i, $i=1,\ldots,n$. The mean value of the x_i's can be computed as follows:

$$\overline{x}=\frac{1}{n}\sum_{i=1}^{n} x_i . \tag{3}$$

For the purpose of robustness, a watermark value should be encoded by moving its mean $\overline{x}$ to the middle of a corresponding quantization interval such that the modulated $\overline{x}$ cannot be easily moved away from the current interval. The mean-quantization-based fragile-watermarking approach operates as follows. Let w be the target watermark symbol to be encoded, and let $\overline{r}$ be the quantization noise defined as

$$\overline{r}=\overline{x}-\left\lfloor \frac{\overline{x}}{q} \right\rfloor \cdot q, \tag{4}$$

where $\lfloor \cdot \rfloor$ is the *floor* operator. To encode w, the amount of update $\overline{u}$ added to the mean coefficient $\overline{x}$ can be determined as follows:

$$\overline{u}=\begin{cases} -\overline{r}+0.5q & \text{if } Q(\overline{x},q)=w, \\ -\overline{r}+1.5q & \text{if } Q(\overline{x},q)\neq w \quad \text{and} \quad \overline{r}>0.5q, \\ -\overline{r}-0.5q & \text{if } Q(\overline{x},q)\neq w \quad \text{and} \quad \overline{r}\leqslant 0.5q. \end{cases} \tag{5}$$

As a consequence, the new mean coefficient becomes $\overline{x}^* = \overline{x}+\overline{u}$. In Eq. (5), $0.5q$ and $1.5q$ are used to shift a mean coefficient $\overline{x}$ to the middle of a quantization interval such that $\overline{x}^*$ is relatively difficult to move away from the current interval. However, updating the mean coefficient implies that all the constituent coefficients need to be updated accordingly by

$$x_i^* = x_i+\overline{u} \ (1\leqslant i\leqslant n), \tag{6}$$

where x_i^* is an updated wavelet coefficient.

Let a modified wavelet coefficient $\hat{x}$ be modeled as

$$\hat{x}=x^*+\Delta, \tag{7}$$

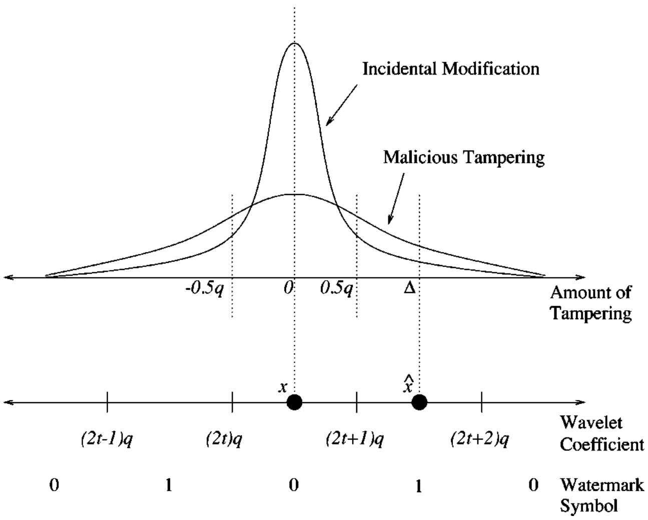


Fig. 1 The statistical distributions of incidental modification and malicious tampering on wavelet coefficients (top), and an illustration of quantization-based watermarking (bottom).

where x^* is the wavelet coefficient defined in Eq. (6) and Δ represents the amount of update caused by tampering. In the case of incidental modification, Δ_I can be modeled as a Gaussian distribution with a smaller variance, that is,

$$\Delta_I \sim N(0,\sigma_I^2), \tag{8}$$

where σ_I denotes the variance of the modification quantities due to an incidental distortion. On the other hand, for an instance of malicious tampering, Δ_M can be modeled as a Gaussian distribution with a larger variance, that is,

$$\Delta_M \sim N(0,\sigma_M^2), \tag{9}$$

where σ_M denotes the variance of modification quantities caused by malicious tampering. Usually, it is assumed that the variance of modification quantities caused by an incidental distortion is smaller than that caused by an instance of malicious tampering, i.e., $\sigma_I < \sigma_M$. Lin and Chang[4] have provided some reference values for σ_I and σ_M in the spatial domain.

Figure 1 illustrates the statistical distributions of updates of wavelet coefficients corresponding to incidental and malicious modifications. In Fig. 1, each quantization interval has a corresponding binary watermark symbol, 0 or 1. The watermark symbol associated with the coefficient x changes when the amount of tampering, Δ, is greater than $|0.5q|$. Based on the above design, a wavelet coefficient is put at the middle of a quantization interval in order to reduce the probability of watermark errors caused by tampering. Since the same watermark symbol appears periodically, the watermark symbol may not be changed even for $\Delta > |0.5q|$. For example, if $\Delta = 2.0q$, the tampered coefficient $\hat{x}$ will fall into the interval $[(2t+2)q,(2t+3)q]$ with the watermark symbol 0, which is the same as the original watermark symbol carried by x. This is the common drawback of a quantization-based watermarking approach. However, since the variance of modification quantities caused by an instance of malicious tampering is larger than that caused by incidental distortion, we can expect that an incidentally distorted coefficient has greater probability of falling into the interval $[-0.5q, 0.5q]$. Thus, we have the hypothesis that the probability of watermark errors caused by an incidental distortion is smaller than that caused by an instance of malicious tampering. In addition, we conduct an analysis in the next paragraph to prove that our scheme can alleviate the well-known drawback of the conventional quantization-based approach.

In order to ensure that an authentication system is incidental-distortion-tolerant, the credibility of a fragile watermark should be increased so that an incidental modification will not be misunderstood as a malicious one. Because the sum of more than two random variables with Gaussian distribution is still a Gaussian distribution but with smaller variance, we have

$$\overline{\Delta} \sim N\left(0, \frac{1}{n}\sigma^2\right) \tag{10}$$

when $\Delta \sim N(0,\sigma^2)$. Thus, when mean quantization is applied, the distribution of modification quantities caused by malicious or incidental distortions will become

$$\overline{\Delta_I} \sim N\left(0, \frac{1}{n}\sigma_I^2\right) \tag{11}$$

and

$$\overline{\Delta_M} \sim N\left(0, \frac{1}{n}\sigma_M^2\right), \tag{12}$$

respectively. Equations (11) and (12) indicate that the proposed mean-quantization-based approach can reduce the variance of modification quantities caused by incidental and malicious distortions, respectively. From Eqs. (11) and (12), it is obvious that when the number of coefficients, n, used to encode a watermark value is increased, the probability of watermark errors will be decreased. In order to increase the credibility of a fragile watermark for image authentication, the watermark errors caused by an instance of malicious tampering should be maximized, and those caused by an incidental distortion should be minimized. Under these circumstances, if the value of n is too small, then the embedded watermark will be too fragile to tolerate incidental manipulation. On the other hand, if n is too large, the embedded watermark will be too robust to detect malicious tampering. Therefore, the number n used to encode a watermark value is a key factor in the trade-off between robustness and fragility. We conduct an analysis with regard to this trade-off in Sec. 2.3.

2.3 *Choosing an Optimal n for Mean Quantization*

In this subsection, we provide a formal proof to show that the proposed mean-quantization-based fragile watermarking scheme is superior to the conventional quantization-based approach.[5] In Ref. 5, Kundur and Hatzinakos assumed that the distributions of modification quantities caused by an instance of malicious tampering and an incidental distortion are both Gaussian. They also mentioned that the major difference between the two distributions is that the variance of the distribution caused by malicious tampering is larger than that caused by incidental distortion. Since the operation of mean quantization will make the variance of all distributions smaller, in this section we devise a systematic way to determine an optimal number of coefficients that should be adopted in the mean quantization process.

Given a distribution of tampering $N(0,\sigma^2)$, and a quantization interval size q, the probability of watermark errors computed using a quantization-based approach is

$$E = 2\sum_{j=0}^{\infty} \int_{(2j+1/2)q}^{(2j+3/2)q} \frac{1}{\sqrt{2\pi}\sigma} \exp - \frac{1}{2}\left(\frac{x}{\sigma}\right)^2 dx \tag{13}$$

$$= 2\sum_{j=0}^{\infty} \lim_{r\to\infty} \sum_{k=0}^{r} \frac{1}{\sqrt{2\pi}\sigma} \frac{q}{r} \times \exp\left(-\frac{1}{2}\left\{\frac{[2j+\frac{1}{2}+(k+\frac{1}{2})/r]q}{\sigma}\right\}^2\right). \tag{14}$$

Since Eq. (13) is not in a discrete format, we use the form shown in Eq. (14) instead to compute the probability of watermark errors with respect to σ and q, because σ and q are two important factors that will influence the results. Figure 2 shows the relations between the variance σ, of tampering, the size q of a quantization interval, and the probability E of watermark errors. The X axis and Y axis in Fig. 2 represent σ/q and E, respectively. However, owing to the fact that the maximum q is bounded by the characteristics of the human visual system,[11] the probability of watermark errors cannot be arbitrarily reduced. On the other hand, for a fixed q, a larger σ value will lead to a larger E value. If the variance σ can be reduced, then the probability of watermark errors caused by a malicious distortion or an incidental distortion will be reduced.

In Eq. (14), we know that the probability of watermark errors is a function of σ/q. Therefore, we can represent the probability by means of $f(t)$, where $t=\sigma/q$. In general, the range of t can be divided into three zones. In the *robust zone* the value of $f(t)$ is very close to 0. In the *fragile zone* the value of $f(t)$ is close to 0.5. There is a transition zone in between, which we call the *semifragile zone*. The value of $f(t)$ changes from 0 to 0.5 within that zone. Therefore, there are two critical points that need to be determined. One is the point at which the value of $f(t)$ changes from zero to nonzero. The other is the point where $f(t)$ starts to saturate at 0.5. We call these points t_1 and t_2, respectively. Furthermore, since the semifragile zone is an ambiguous zone, we would like to make it as small as possible. For this purpose, the values of t_1 and t_2 can be determined by solving the following constraint optimization problem:

$$g(t_1,t_2) = \alpha\cdot|f(t_1)| + \beta\cdot|f(t_2)-0.5| + \gamma\cdot\frac{t_2}{t_1}, \tag{15}$$

where $g(\cdot)$ is a cost function to be minimized. The first term and the second term on the right-hand side of Eq. (15) are the constraints that force the values of $f(t_1)$ and $f(t_2)$ to be as close as possible to 0 and 0.5, respectively. As to

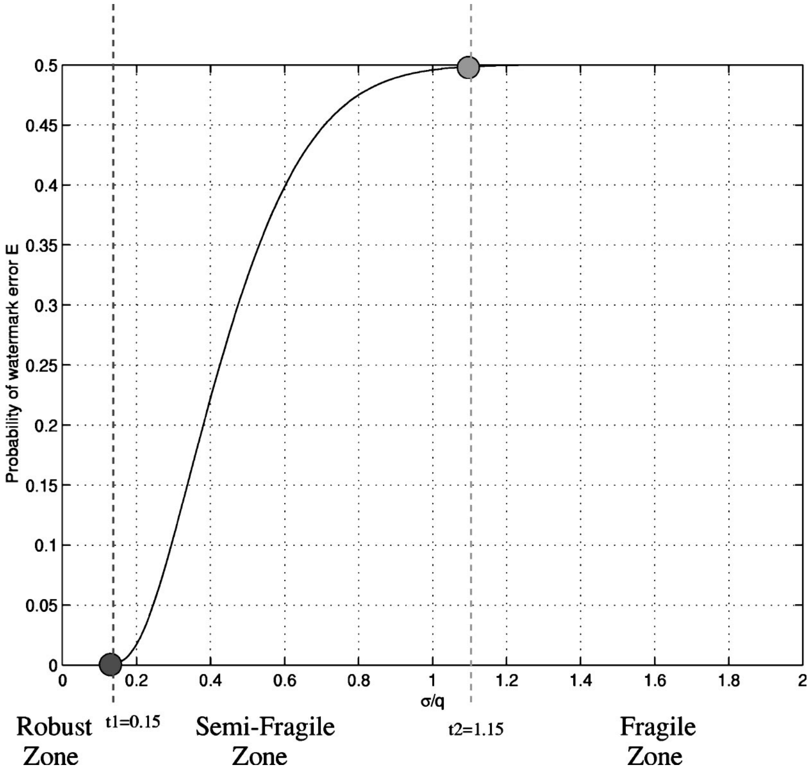


Fig. 2 The relation between the variance of tampering σ, the quantization interval's size q, and the probability of watermark errors.

the t_2/t_1 term, it is used to keep the size of the transition zone as small as possible. On the other hand, the parameters α, β, and γ are designed to control the values of t_1 and t_2, which will determine the ranges of the three zones shown in Fig. 2. In fact, t_1 and t_2 together can determine the number of coefficients (n) needed to embed a watermark. Therefore, the selection of α, β, and γ will certainly influence the selection of n. However, since the best n can be chosen by optimizing an objective function, obtaining different optimal n is possible if different sets of α, β, and γ are chosen. In this paper, the values of α and β should be set the same because their importance is equal. On the other hand, the relation $\alpha=\beta\gg\gamma$ should hold, so that the boundaries of the three zones will be clear-cuts. One thing to be noted is that when that relation holds, different sets of α, β, and γ will not influence the value of n. In our experiments, we set the values of the leading coefficients α, β, and γ at 1000, 1000, and 1, respectively. Based on the above setting, t_1 and t_2 can be determined. They are 0.15 and 1.15, respectively.

Let the distribution of an instance of malicious tampering and an incidental distortion be denoted as $N(0,\sigma_I^2)$ and $N(0,\sigma_M^2)$, respectively. From Lin and Chang's[4] previous experience, we know that σ_M is larger than σ_I, and they have a relation $\sigma_M=c\sigma_I$ with $c>1$. Let n denote the number of coefficients used in calculating a mean coefficient [Eq. (3)]; the new distributions of modification quantities caused by a malicious tampering and an incidental distortion become $N(0,(\sigma_I^*)^2)$ and $N(0,(\sigma_M^*)^2)$, respectively, where $\sigma_I^*=(1/\sqrt{n})\sigma_I$ and $\sigma_M^*=(1/\sqrt{n})\sigma_M=(c/\sqrt{n})\sigma_I$. Let the size of a quantization interval, q, be determined according to the human visual system.[11] This means that q is fixed with respect to the human visual system. The question is how to determine the best n such that the probability of watermark errors caused by an instance of malicious tampering will be maximized and that caused by an incidental distortion will be minimized. If the relation $\sigma_M^*/q \geqslant t_2$ holds, then the probability of watermark errors caused by a malicious tampering will definitely be maximized. Therefore, we have

$$\frac{\sigma_M^*}{q}\geqslant t_2 \Rightarrow \frac{c\sigma_I}{\sqrt{n}q}\geqslant t_2 \Rightarrow \frac{c\sigma_I}{t_2 q}\geqslant\sqrt{n}. \tag{16}$$

Similarly, if the relation $\sigma_I^*/q\leqslant t_1$ holds, then the probability of watermark errors caused by an incidental distortion will be minimized. That is,

$$\frac{\sigma_I^*}{q}\leqslant t_1 \Rightarrow \frac{\sigma_I}{\sqrt{n}q}\leqslant t_1 \Rightarrow \frac{\sigma_I}{t_1 q}\leqslant\sqrt{n}. \tag{17}$$

Combining Eqs. (16) and (17), we obtain

$$\frac{\sigma_I}{t_1 q}\leqslant\sqrt{n}\leqslant\frac{c\sigma_I}{t_2 q}. \tag{18}$$

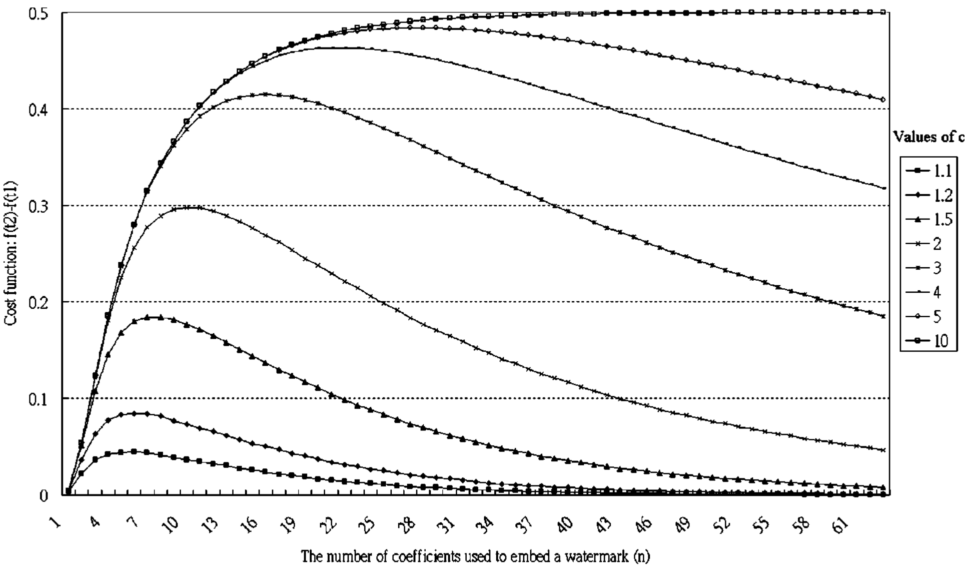


Fig. 3 The objective function, $f(t_2)-f(t_1)$, as a function of c and n.

It is obvious that the minimum n that satisfies Eq. (18) is an n_1 that makes $\sigma_I/t_1 q=\sqrt{n_1}$. Therefore, we have

$$n_1=\left(\frac{\sigma_I}{t_1 q}\right)^2. \tag{19}$$

This n_1 will lead to the minimum probability of watermark errors caused by an incidental distortion. On the other hand, the maximum n that will satisfy Eq. (18) is an n_2 that makes $\sqrt{n_2}=c\sigma_I/t_2 q$. Thus, we have

$$n_2=\left(\frac{c\sigma_I}{t_2 q}\right)^2. \tag{20}$$

This n_2 will lead to the maximum probability of watermark errors caused by an instance of malicious tampering. In order to find the best n that will bypass an incidental distortion while detecting an instance of malicious tampering, we should select an n that is bounded by n_1 and n_2, i.e., $n\in[n_1,n_2]$.

In what follows, we shall conduct a theoretical analysis to determine an ideal n. From Eq. (14), we know that the probability of watermark errors is a function of σ/q. Since q is a constant when a specific human visual model[11] is adopted, t is proportional to σ. Let the probabilities of watermark errors caused by an incidental distortion and a malicious tampering be $f(\hat{t}_1)$ and $f(\hat{t}_2)$, respectively, where $\hat{t}_1=\sigma_I/q$ and $\hat{t}_2=\sigma_M/q$. Because $\sigma_M=c\sigma_I$, we have

$$\hat{t}_2=\frac{\sigma_M}{q}=\frac{c\sigma_I}{q}=c\hat{t}_1. \tag{21}$$

When a mean-quantization operation covering n coefficients is applied, σ_I and σ_M will be updated to $\sigma_I/\sqrt{n}$ and $\sigma_M/\sqrt{n}$, respectively. In order to obtain the best mean-quantization result, the difference between the watermark errors caused by an instance of malicious tampering and an incidental distortion should be maximized. That is, $f(\hat{t}_2)-f(\hat{t}_1)$ should be maximized. The physical meaning of maximizing $f(\hat{t}_2)-f(\hat{t}_1)$ is to make the watermark errors caused by an instance of malicious tampering as large as possible and those caused by an incidental distortion as small as possible. Using the optimization scheme, one can decide on an optimal value of n such that $f(\hat{t}_2)-f(\hat{t}_1)$ is maximized. The simplest way to calculate the ideal n is to compute the values of $f(c\sigma_I/\sqrt{n}q)-f(\sigma_I/\sqrt{n}q)$ using various integers $n\in[n_1,n_2]$. The integer that leads to the largest outcome is the ideal n. Figure 3 shows a 2-D plot of the maximization function, $f(\hat{t}_2)-f(\hat{t}_1)$, as a function of c and n.

3 Tampered-Area Estimation Using Information Fusion

For image authentication, the wavelet-based fragile-watermarking method proposed in Ref. 5 only shows the tampering detection results at multiple scales. In this section, we present an information fusion technique that can be used to integrate the results obtained at multiple scales. In addition, the proposed technique has the merit of suppressing sparse watermark errors spread out over the subimages at multiple scales.

The analysis conducted in Sec. 2.3 provides a procedure to compute the optimal n of every subband at different scales. Due to the length of a quantization interval, the variance of distribution of an instance of malicious tampering and that of an incidental distortion are different in different subbands; the optimal n's of the LH, HL, and HH subbands may be different. Let the size of an image be $N\times N$, where N is a power of 2. We embed the watermark by changing the coefficients of one of the LH, HL, and HH subbands. Since the optimal n at each subband may be different, we arrange the ordering of a sequence so that the embedded watermark is secure, robust, and localized. In what follows, we first discuss the operation at a scale without specifying the scale until we need to integrate the results. Let the optimal *n* of the LH, HL, and HH subbands at scale *l* be denoted as n_{LH}, n_{HL}, and n_{HH}, respectively. In addition, we rename and reorder n_{LH}, n_{HL}, and n_{HH} as $n_1 \leqslant n_2 \leqslant n_3$. At scale *l*, the size of every subband becomes $L \times L$, where $L = N/2^l$. On the other hand, we map every coefficient $x(i,j)$ at position (i,j) into a 1-D value $y(p)$ using the following formula:

$$p = \left\lfloor \frac{i}{w} \right\rfloor \cdot wL + hw + o, \tag{22}$$

where $w = \lceil \sqrt{n_2} \rceil$,

$$h = \begin{cases} j & \text{if } \lfloor i/w \rfloor \text{ is even,} \\ L-j & \text{if } \lfloor i/w \rfloor \text{ is odd,} \end{cases}$$

and $o = i - \lfloor i/w \rfloor$. From the above transformation, we establish a one-to-one relation between the 2-D representation $x(i,j)$ and the 1-D representation $y(p)$. Before we embed a watermark in a specific subband, a sequence s_k is first generated by a private key, where $s_k \in \{\mathrm{LH,HL,HH}\}$. The value of s_k will indicate in which subband the *k*'th watermark value should embed. We denote the number of coefficients used to embed the *k*'th watermark as n_{s_k}. The coefficients, $C_k = \{y(p^*), y(p^*+1), \ldots, y(p^*+n_{s_k}-1)\}$, in the subband s_k are allocated for embedding the *k*'th watermark, where $p^* = \Sigma_{i=1}^{k-1} n_{s_i}$. Finally, the *k*'th watermark will be embedded in the set of coefficients C_k using the mean-quantization embedding rule described in Eq. (6).

In what follows we compute the probability of watermark errors caused, respectively, by malicious tampering and incidental distortion. Based on the analysis of these probabilities, we are able to judge what regions have been maliciously tampered with. A set of coefficients, C_j, is defined as a *neighbor* of C_i if any coefficient in C_j is four-connected to C_i. We denote the set of neighbors of C_k as

$$S_k = \{C_i \mid C_i \text{ is a neighbor of } C_k\}.$$

Let T_k denote the status of a malicious tampering corresponding to the coefficient C_k, i.e.,

$$T_k = \begin{cases} 1 & \text{if } C_k \text{ has watermark error,} \\ 0 & \text{otherwise.} \end{cases}$$

Let the theoretical probability of watermark errors caused by an instance of malicious tampering in subband *s* be denoted as P_s^M, and that caused by an incidental distortion be denoted as P_s^I, where $s \in \{\mathrm{LH,HL,HH}\}$. The value of $P_s^M - P_s^I$ can be maximized by choosing the optimal *n* by using the method described in Sec. 2.3. Under these circumstances, the estimated probability that C_k has been maliciously tampered with is

$$E_k = \frac{M_k}{M_k + I_k}, \tag{23}$$

where

$$M_k = \prod_{\{k^* \mid C_{k^*} \in (S_k \cup \{C_k\})\}} [T_{k^*} P^M_{s_{k^*}} + (1-T_{k^*})(1-P^M_{s_{k^*}})] \tag{24}$$

indicates the probability that the watermark error detected in $S_k \cup \{C_k\}$ is caused by an instance of malicious tampering, and

$$I_k = \prod_{\{k^* \mid C_{k^*} \in (S_k \cup \{C_k\})\}} [T_{k^*} P^I_{s_{k^*}} + (1-T_{k^*})(1-P^I_{s_{k^*}})] \tag{25}$$

indicates the probability that the detected watermark error is caused by an incidental distortion. E_k here indicates the probability that the coefficients in C_k have been maliciously tampered with. The set of coefficients in C_k is

$$C_k = \{y(p_1), y(p_2), \ldots, y(p_{n_{s_k}})\}$$
$$= \{x(i_1, j_1), x(i_2, j_2), \ldots, x(i_{n_{s_k}}, j_{n_{s_k}})\}.$$

The relation between p_k and (i_k, j_k) has been given in Eq. (22). In order to specify it at the scale level, we use E_k^l to represent, at scale *l*, the probability that the coefficients in C_k are maliciously tampered with. Under these circumstances, for a specific position (i^0, j^0) at scale *l*, the probability of its being tampered with can be computed by

$$E^l_{i^0,j^0} = E^l_{k^*}, \tag{26}$$

where $x(\lfloor i^0/2^l \rfloor, \lfloor j^0/2^l \rfloor) \in C_{k^*}$. The probability of watermark errors caused by an instance of malicious tampering can thus be computed by integrating the information detected at each scale using the following rule;

$$E_{i^0,j^0} = \prod_{l=1}^{\mathrm{Scale}} E^l_{i^0,j^0}, \tag{27}$$

where Scale represents the number of scales used in the wavelet transform. In order to detect the complete tampered area, we extract the areas where the probability is higher as our final results. For achieving this goal, we use a rule as follows: If a pixel at position (*i*,*j*) is maliciously tampered with, then

$$E_{i,j} \geqslant (0.5)^{\mathrm{Scale}}. \tag{28}$$

The threshold used in Eq. (28) is learned by experience. But the underlying assumption is that M_k should be larger than I_k for a malicious tampering.

The mentioned mechanism can be used to detect most of the areas that have been maliciously tampered with. However, when such an area is very small, it is difficult to distinguish it from an area that has encountered incidental distortion. This is because an instance of malicious tamper-

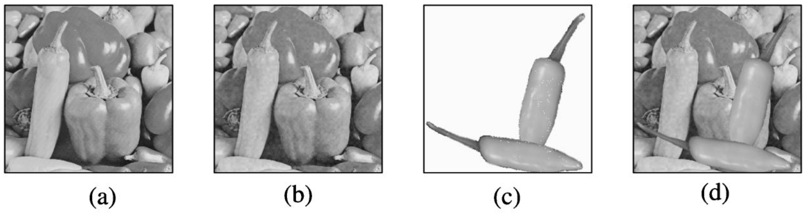


Fig. 4 An example showing malicious tampering by means of object replacement: (a) original image; (b) watermarked image; (c) objects used for tampering; (d) modified watermarked image.

ing and an incidental distortion both generate watermark errors of the sparse type. However, these small watermark errors will collapse if the evidence located at different scales are integrated. On the other hand, if the probability of watermark errors caused by an incidental distortion is very small (zero is the ideal case), then one can claim that the detected watermark errors were completely obtained from an area that was maliciously tampered with.

4 Experimental Results

To demonstrate the power of our image authentication system, we first introduce the experimental setup in Sec. 4.1 and then give the detection results obtained under various incidental distortions in Sec. 4.2. In Sec. 4.3, we present some experimental results obtained by applying both malicious tampering and incidental manipulation. A set of test images processed by combining different incidental and malicious manipulations was used to estimate the area that was maliciously tampered with. A comparison of the performance of the conventional quantization-based approach and our approach will be made in Sec. 4.4.

4.1 *Experimental Setup*

The images used in the experiment were of size 512×512 with 256 gray levels. Figure 4 is an example showing how a watermarked image is tampered with, including the original image, the watermarked image, the altered area, and the final altered image. The PSNR of the watermarked image shown in Fig. 4(b) was 35.91 dB. Two peppers [Fig. 4(c)] were added as shown in Fig. 4(b) and formed an image that had been tampered with, as shown in Fig. 4(d). This set of data was used to test the performance of our approach in the subsequent experiments.

The set of incidental attacks used in the experiments included JPEG compression, blurring, and sharpening. The mask sizes used in the blurring operation were 3×3, 5×5, and 7×7, respectively. The quality factors adopted for JPEG compression were from 10% to 90%, and the parameters used in the sharpening operation were from 10% to 50%. In the experiments, the watermark sequence was embedded in one of the LH, HL, and HH subbands randomly at each scale of a wavelet-transformed image. As to the determination of the best n at every scale of a wavelet transform, this can be calculated by scanning the interval $[n_1, n_2]$ for large c ($c > 7.67$) or by scanning the interval $[n_2, n_1]$ for small c ($1 < c \leq 7.67$), where n_1 and n_2 are computed using Eqs. (19) and (20), respectively. We use $\mathbf{n} = (n^1_{\mathrm{LH}}, n^1_{\mathrm{HL}}, n^1_{\mathrm{HH}}; n^2_{\mathrm{LH}}, n^2_{\mathrm{HL}}, n^2_{\mathrm{HH}}; \ldots; n^s_{\mathrm{LH}}, n^s_{\mathrm{HL}}, n^s_{\mathrm{HH}})$ to represent the number of coefficients used at every scale in the mean quantization process, where n_i is the number of coefficients used to derive a mean at scale i, and s is the total number of scales used. From Eqs. (19) and (20), the best set of n could be theoretically determined as (9, 9, 7; 16, 16, 12; 16, 16, 13; 11, 11, 8) when the total number of scales was chosen to be 4.

4.2 *Detection Results Obtained by Applying Incidental Distortions Only*

In this section, we check whether our approach could tolerate a number of incidental operations with different de-

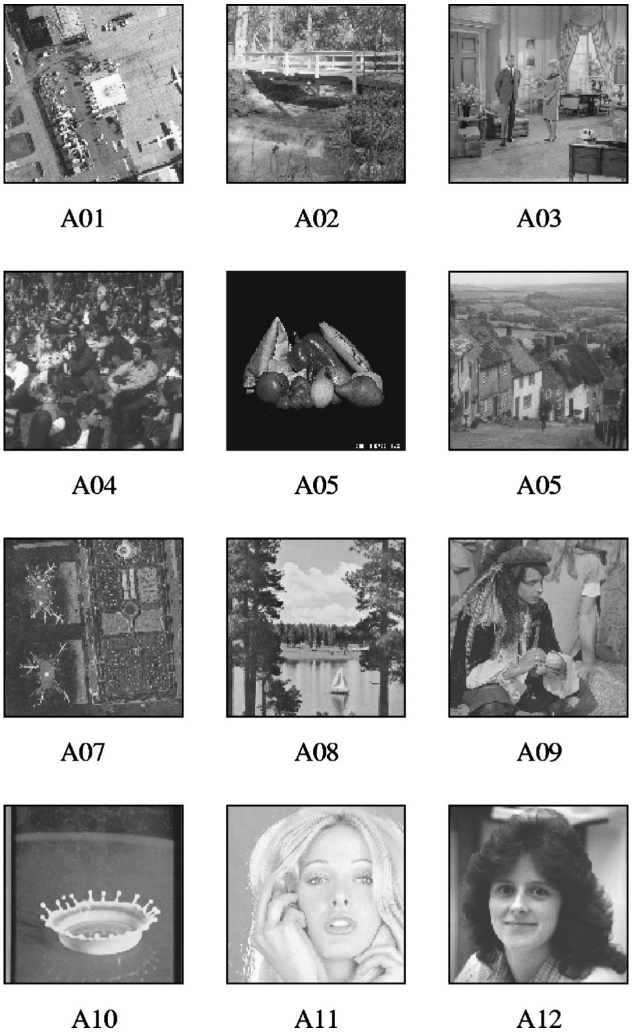


Fig. 5 A set of test images.

Table 1 Tampering detection for a set of incidentally manipulated test images. A √ symbol indicates that our system treats the operation as an incidental distortion, while a × symbol indicates that the operation was misidentified as malicious tampering.

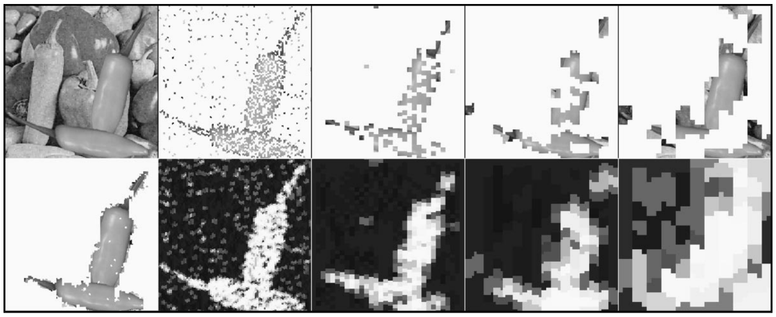
Image	Response											
Operation	Image A01	A02	A03	A04	A05	A06	A07	A08	A09	A10	A11	A12
Blur (3×3)	×	×	×	×	×	×	×	×	×	×	×	√
Blur (5×5)	×	×	×	×	×	×	×	×	×	×	×	×
Blur (7×7)	×	×	×	×	×	×	×	×	×	×	×	×
Sharpen (*F*=10%)	√	√	√	√	√	√	√	√	√	√	√	√
Sharpen (*F*=20%)	√	√	√	√	√	√	√	√	√	√	√	√
Sharpen (*F*=30%)	√	√	√	√	√	√	√	√	√	√	√	√
Sharpen (*F*=40%)	√	×	√	√	√	√	√	×	√	√	√	√
Sharpen (*F*=50%)	√	×	×	×	√	×	×	×	√	×	×	√
Sharpen (*F*=60%)	×	×	×	×	×	×	×	×	×	×	×	×
Sharpen (*F*=70%)	×	×	×	×	×	×	×	×	×	×	×	×
Sharpen (*F*=80%)	×	×	×	×	×	×	×	×	×	×	×	×
Sharpen (*F*=90%)	×	×	×	×	×	×	×	×	×	×	×	×
JPEG (QF=90%)	√	√	√	√	√	√	√	√	√	√	√	√
JPEG (QF=80%)	√	√	√	√	√	√	√	√	√	√	√	√
JPEG (QF=70%)	√	√	√	√	×	√	√	√	√	√	√	√
JPEG (QF=60%)	√	√	√	√	×	√	√	√	√	√	√	√
JPEG (QF=50%)	√	√	√	√	×	√	√	√	√	√	√	√
JPEG (QF=40%)	√	√	√	√	×	√	√	√	√	√	√	√
JPEG (QF=30%)	√	√	√	√	×	√	√	√	√	√	√	√
JPEG (QF=20%)	√	×	×	×	×	×	×	×	×	×	×	×
JPEG (QF=10%)	×	×	×	×	×	×	×	×	×	×	×	×

grees of alteration. Figure 5 shows a set of test images that was used in the experiments. The incidental operations that were applied to the set of test images included JPEG compression, blurring, and sharpening. Table 1 lists the results obtained in this experiment. A √ symbol indicates that our system considered the operation to be an incidental one. On the other hand, a × symbol indicates that our system mistakenly considered the operation to be a malicious one. From the table, it is obvious that our system could successfully pass almost all the JPEG-compressed images down to quality factor 30%. As for the sharpening operation, our system could successfully tolerate most of the sharpened images up to a 40% sharpening factor. However, in the case of the blurring operation, our system did not work well.

4.3 *Detection Results Obtained by Applying Malicious Tampering and Incidental Manipulation Simultaneously*

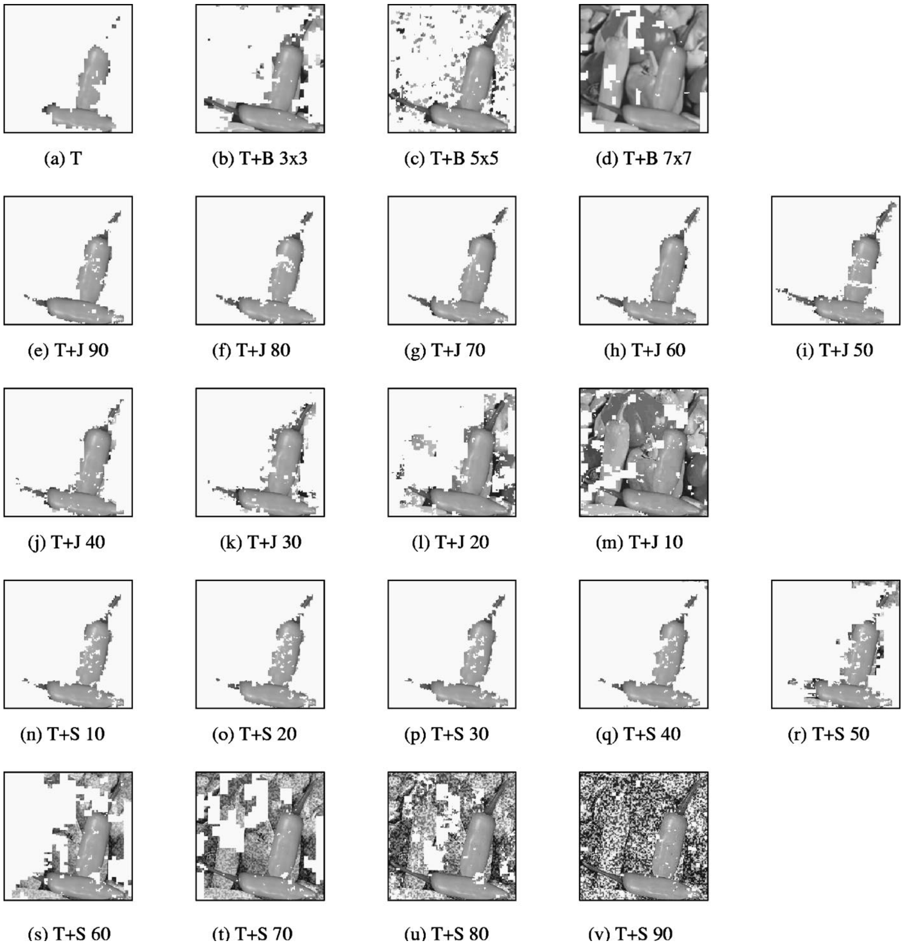
In this section, we give some experimental results obtained by applying malicious tampering and an incidental manipulation simultaneously. The objective of these experiments was to check whether our approach could successfully tolerate an incidental manipulation while detecting a malicious attack. Figure 6(a) is a pepper image that was modified by performing 60% (quality factor) JPEG compression, followed by two-pepper replacement. The detected watermark errors at scales 1 to 4 are shown in Figs. 6(b) to 6(e), respectively. It can be seen that the watermark errors caused by the JPEG compression are much fewer than those caused by malicious tampering. The detected watermark errors were then converted into the probability of been maliciously tampered with as shown in Figs. 6(g) and 6(j). It is obvious that the coefficients having the sparse type all had lower probability of having been maliciously tampered with at each scale. On the other hand, the areas that corresponded to the regions that were maliciously tampered with all had higher probability of having been maliciously tampered with. After performing information fusion, the final detected altered areas were those shown in Fig. 6(f). It is apparent that the maliciously modified regions were detected correctly.

Figure 7 shows another 21 detection results obtained using the proposed mean-quantization-based fragile-watermarking technique. The symbols T, B, J, and S denote malicious tampering, blurring, JPEG compression, and sharpening, respectively. The number following each symbol is the parameter used in an incidental distortion. For example, ''T+B 3×3'' in Fig. 7(b) means an image was maliciously tampered with and then blurred with a mask of size 3×3. In the whole set of experiments, the resolution of the wavelet transform was taken up to 4 scales. The optimal number of coefficients used to perform mean quantization at each scale was $\mathbf{n}=(9,9,7;16,16,12;16,16,13;11,11,8)$. From Fig. 7, it is apparent that our approach did work well in most cases, especially in tolerating incidental manipulation like JPEG. Figure 7(l) indicates that when the quality



(a)	(b)	(c)	(d)	(e)
(f)	(g)	(h)	(i)	(j)

Fig. 6 Tampering with object placement and JPEG compression: (a) is a tampered image with two objects added; (b) to (e) are the detected watermark errors from scales 1 to 4, respectively; (g) to (j) are the tamper response maps derived from scales 1 to 4, respectively; (f) is the final result after performing information fusion.



(a) T (b) T+B 3x3 (c) T+B 5x5 (d) T+B 7x7

(e) T+J 90 (f) T+J 80 (g) T+J 70 (h) T+J 60 (i) T+J 50

(j) T+J 40 (k) T+J 30 (l) T+J 20 (m) T+J 10

(n) T+S 10 (o) T+S 20 (p) T+S 30 (q) T+S 40 (r) T+S 50

(s) T+S 60 (t) T+S 70 (u) T+S 80 (v) T+S 90

Fig. 7 A set of detection results obtained by applying our mean-quantization-based method. (a) is the detection result when the attack is object placement only; (b) to (d) show the detection results when the attack is object placement followed by blurring with mask sizes of 3×3, 5×5, and 7×7, respectively; (e) to (m) show the detected results when the attack is object placement followed by JPEG compression with a quality factor ranging from 90% to 10% in steps of 10%; (n) to (v) show the detection results when the attack is object placement followed by sharpening with a sharpening factor ranging from 10% to 90% in steps of 10%.

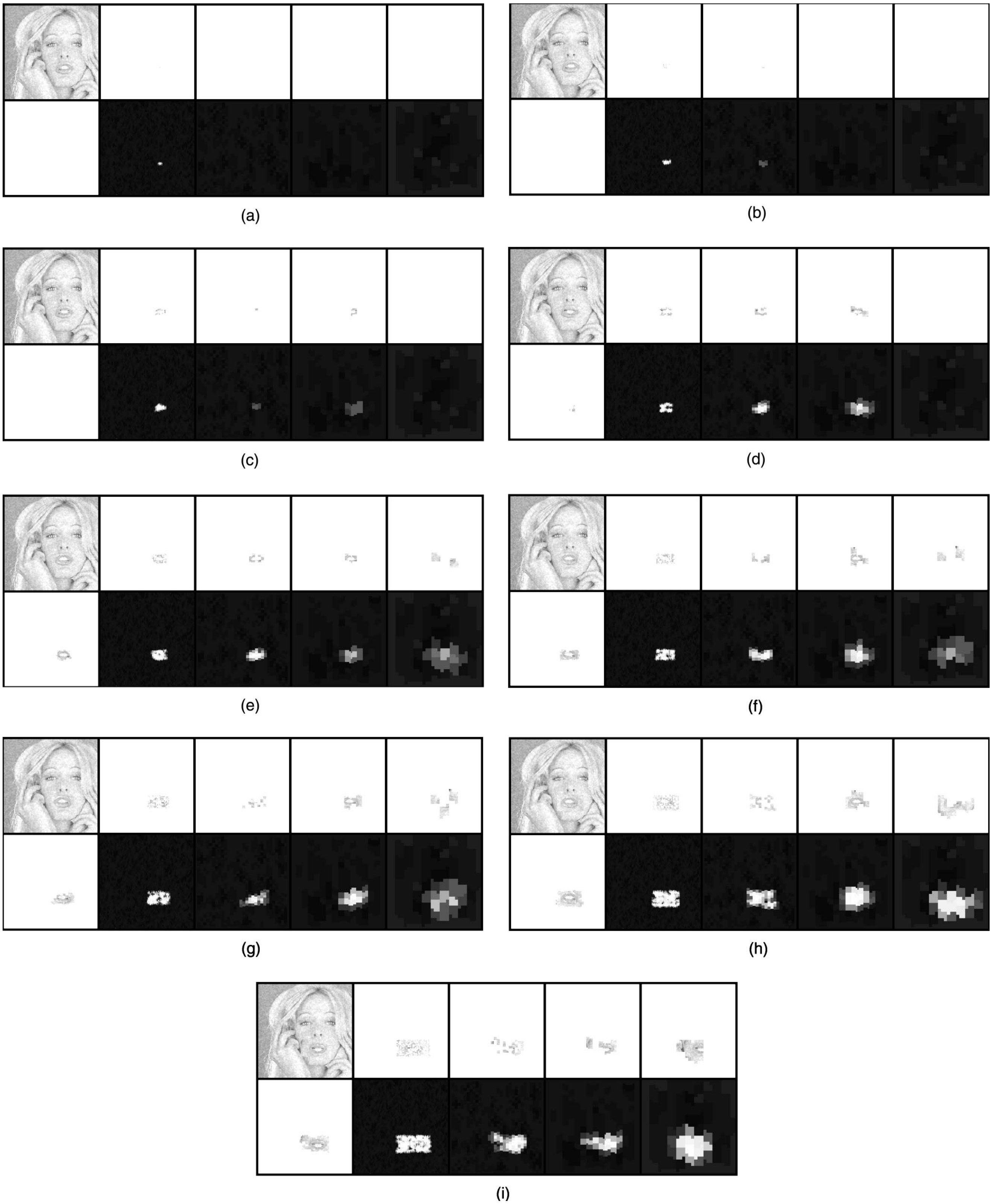


Fig. 8 Sensitivity test of our algorithm against small image modifications.

factor reached 30%, the detection result was still good. In the case of a combined attack including 5×5 and 7×7 blurring [Fig. 7(d)], the results were bad. But when the window size was 3×3, the detection result was good. In the case of a combined attack involving sharpening, the results were good when the sharpening factor was smaller than 40%. When the sharpening factor reached or exceeded 60%, the detected results were completely wrong.

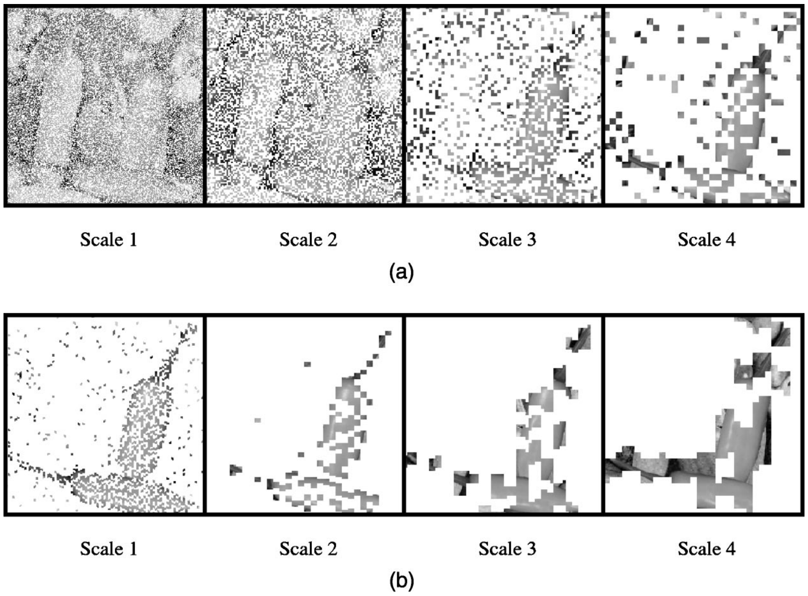


Fig. 9 Comparison of detected watermark errors obtained using the conventional quantization-based approach and the mean-quantization-based approach with $\mathbf{n}=(9,9,7;16,16,12;16,16,13;11,11,8)$.

We also conducted a series of experiments to test the sensitivity of our algorithm to small image modifications. The test image used was the A11 image as shown in Fig. 5. We selected the area close to the mouth as the target to tamper with. We gradually enlarged the tampered areas and the detected results are shown in Figs. 8(a) to 8(i). It is obvious that when the modified area was very small, our algorithm could not detect the change [(a) to (d)]. However, when the modified area reached a certain size, our algorithm was able to detect it correctly [(e) to (i)].

4.4 *Comparison with the Conventional Quantization-based Approach*

In this subsection we compare our approach with the conventional approach. The maliciously attacked image shown in Fig. 4(d), subjected to JPEG compression with a quality factor 60%, was used as the test image. The watermark errors (at scales 1 to 4) obtained by applying the conventional quantization-based approach[5] and the proposed mean-quantization-based approach with $\mathbf{n}=(9,9,7;16,16,12;16,16,13;11,11,8)$ are shown in Figs. 9(a) and 9(b), respectively. It is obvious that the results obtained by applying our approach are better than those obtained by applying the conventional approach.

5 Conclusion

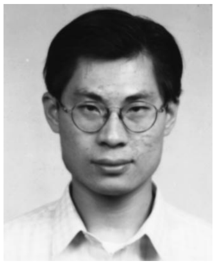
In this paper, a mean-quantization-based fragile-watermarking approach has been proposed for image authentication. Our system is able to maximize the probability of watermark errors caused by an instance of malicious tampering and minimize the probability of watermark errors caused by an incidental distortion. In addition, an information fusion procedure that can integrate detection responses at each scale in the wavelet domain has been presented, which can be used to estimate the area that has been maliciously tampered with.

Our future work will proceed in two directions. First, the capability of our image authentication system in distinguishing malicious tampering and incidental distortion will be further improved so that incidental distortion with large variance of modification, such as histogram equalization, can also be tolerated. Secondly, we will extend the mean-quantization-based watermarking approach to multipurpose watermarking, so that an embedded watermark can be used in multiple applications.

References

1. S. Bhattacharjee and M. Kutter, ‘‘Compression tolerant image authentication,’’ in *IEEE Int. Conf. on Image Processing*, Vol. 1, pp. 4–7 (1998).
2. J. Dittmann, A. Steinmetz, and R. Steinmetz, ‘‘Content-based digital signature for motion pictures authentication and content-fragile watermarking,’’ in *IEEE Int. Conf. Multimedia Computing and Systems*, Vol. II, 1999.
3. G. L. Friedman, ‘‘The trustworthy digital camera: restoring credibility to the photographic image,’’ *IEEE Trans. Consum. Electron.* **39**, 905–910 (1993).
4. C.-Y. Lin and S.-F. Chang, ‘‘A robust image authentication method surviving JPEG lossy compression,’’ in *Int. Conf. on Storage and Retrieval of Image/Video Database, Proc. SPIE* **3312** (1998).
5. D. Kundur and D. Hatzinakos, ‘‘Digital watermarking for telltale tamper proofing and authentication,’’ *Proc. IEEE* **87**, 1167–1180 (1999).
6. R. B. Wolfgang and E. J. Delp, ‘‘Fragile watermarking using the VM2D watermark,’’ in *Int. Conf. on Security and Watermarking of Multimedia Contents*, *Proc. SPIE* **3657**, 204–213 (1999).
7. P. W. Wong, ‘‘A public key watermark for image verification and authentication,’’ in *IEEE Int. Conf. on Image Processing* (1998).
8. M. Wu and B. Liu, ‘‘Watermarking for image authentication,’’ in *IEEE Int. Conf. on Image Processing* (1998).
9. M. Yeung and F. Mintzer, ‘‘An invisible watermarking technique for image verification,’’ in *IEEE Int. Conf. on Image Processing* (1997).
10. B. Zhu, M. D. Swanson, and A. H. Tewfik, ‘‘Transparent robust authentication and distoration measurement technique for images,’’ in *The 7th IEEE Digital Signal Processing Workshop*, pp. 45–48, (1996).
11. A. B. Warson, G. Y. Yang, J. A. Solomon, and J. Villasenor, ‘‘Visibility of wavelet quantization noise,’’ *IEEE Trans. Image Process.* **6**(8), 1164–1175 (1997).

Gwo-Jong Yu received the BS degree in information computer engineering from the Chung-Yuan Christian University, Chung-Li, Taiwan in 1989. Since July 2000, he has been a research assistant in the Institute of Information Science, Academia Sinica, Taipei, Taiwan. He is currently working toward the PhD degree in computer science from National Central University, Chung-Li, Taiwan. His research interests include digital watermarking, face recognition, and neural networks.



Chung-Shien Lu received the PhD degree in electrical engineering from National Cheng-Kung University, Tainan, Taiwan, in 1998. His thesis is about wavelet-based 2-D/3-D texture analysis. From September 1994 to June 1998 he was also a research assistant in the Institute of Information Science, Academia Sinica, Taipei, Taiwan, ROC. Since October 1998, he has been a postdoctoral fellow in the same institute. He was the recipient of the excellent paper award of the Image Processing and Pattern Recognition Society of Taiwan in 2000 for his work on digital watermarking. His current research interests include digital watermarking and data hiding, multimedia signal processing, image processing, and visual communications. He is a member of the IEEE.

Hong-Yuan Mark Liao received the BS degree in physics from National Tsing-Hua University, Hsin-Chu, Taiwan, in 1981, and the MS and PhD degrees in electrical engineering from Northwestern University, Illinois, in 1985 and 1990, respectively. He was a research associate in the Computer Vision and Image Processing Laboratory at Northwestern University during 1990–1991. In July 1991, he joined the Institute of Information Science, Academia Sinica, Taiwan, as an assistant research fellow. He was promoted to associate research fellow and then research fellow in 1995 and 1998. From August 1997 to July 2000, he served as the deputy director of the institute. Currently, he is the acting director of the Institute of Applied Science and Engineering Research, Academia Sinica, Taiwan. Dr. Liao's current research interests include multimedia signal processing, wavelet-based image analysis, content-based multimedia retrieval, and multimedia protection. He was the recipient of the Young Investigators' award of Academia Sinica in 1998, the excellent paper award of the Image Processing and Pattern Recognition Society of Taiwan in 1998 and 2000, and the paper award of the above society in 1996 and 1999. Dr. Liao served as the program chair of the International Symposium on Multimedia Information Processing (ISMIP'1997) and will serve as the program cochair of the second IEEE Pacific-Rim Conference on Multimedia. He also served on the program committees of several international and local conferences. Dr. Liao is on the editorial boards of the *IEEE Transactions on Multimedia,* the *International Journal of Visual Communication and Image Representation, Acta Automatica Sinica,* the *Tamkang Journal of Science and Engineering,* and the *Journal of Information Science and Engineering.* He is a member of the IEEE Computer Society.