MEAN QUANTIZATION BLIND WATERMARKING FOR IMAGE AUTHENTICATION

Gwo-Jong Yu, Chun-Shien Lu**, Hong-Yuan Mark Liao**, and Jang-Ping Sheu**

* Department of Computer Science and Information Engineering
National Central University, Chung-Li, Taiwan
** Institute of Information Science
Academia Sinica, Taipei, Taiwan
e-mail: {yugj, lcs, liao}@iis.sinica.edu.tw, sheujp@axp2.csie.ncu.edu.tw

ABSTRACT

The objective of this paper is to propose an image authentication scheme, which is able to detect malicious tampering of images even they have also been incidentally distorted. By modeling incidental and malicious distortions as Gaussian distributions with small and large variances, respectively, we propose to embed a watermark in the wavelet domain by a mean quantization technique. Due to the various probabilities of tamper response at each scale, these responses are integrated to make a decision on the tampered areas. Statistical analysis is conducted and experimental results are given to demonstrate that our watermarking scheme is able to detect malicious attacks while tolerating incidental distortions.

1. INTRODUCTION

Image authentication becomes very important due to the availability of Internet. To save bandwidth and storage, digital images are usually transmitted or stored in a compressed form. In addition, images may be processed by blurring or equalization operations by users for specific purposes. Thus, an image authentication system should be able to tolerate incidental modifications while detecting malicious updates. A number of researches used digital signatures for image authentication [1, 2, 3, 5]. Bhattacharjee and Kutter [1] extracted salient feature points and store their positions as the digital signature. Because feature points are assumed not to be shifted too much under incidental distortion, the tampered area can be identified as those area where their corresponding feature points mismatch. Lin and Chang [5] stored the relation of DCT coefficients at all pairs of two random 8×8 blocks as digital signature. They proved that these relations are invariant to JPEG compression. Their method can detect malicious tampering under JPEG compression. The main disadvantage of digital signature-based methods is that it cannot be used for multipurpose watermarking [6] since the image is not watermarked. On the other hand, the watermark-based image authentication approaches detect tampering based on the fragility of the hidden watermark. Kundur and Hatzinakos [4] embedded a watermark value by modulating a selected wavelet coefficient into the quantized interval determined from the corresponding watermark value. They defined that the type of tampering is JPEG compression if the TAF values decrease monotonically from high resolution to low resolution. However, they didn't provide a mechanism to detect the combination of malicious tampering and incidental distortion. Recently, Lu et al. [6, 7] proposed multipurpose watermarking scheme for image and audio authentication and protection. They combined an asymmetric quantization technique and complementary watermarks [8] to achieve a certain degree of robustness and fragility. Although the survival of incidental manipulations and the stability have been improved, their methods still fail to resist some incidental distortions (like Gaussian noise adding).

Because quantization-based watermarking is very sensitive to modification, it is necessary to decrease the fragility of the hidden watermark in order to distinguish malicious tampering from incidental distortion. Based on the observation of Lin and Chang [5], increase of robustness will not sacrifice fragility too much because individual distortion always has smaller variance. The robustness of an embedded watermark can be improved by either enlarging the quantization interval or reducing the amount of modification caused by image processing. Given a set of samples, the population mean has a smaller variance than that of individual samples. It is expected that watermark embedded by modulating the mean of wavelet coefficients is more robust than by modulating individual coefficient. It has been observed that the changes of wavelet coefficients have the tendency of increasing magnitudes under equalization or sharpening and decreasing magnitudes under JPEG compression or blurring[8]. To make the embedded watermark robust to incidental distortions, the watermark is embedded by modulating the mean value of weighted magnitudes of wavelet coefficients. The amount of modification on wavelet coefficients can be modeled as Gaussian distribution with small and large variances for incidental distortion and malicious tampering, respectively. Thus, the probability of watermark error caused by incidental distortion is smaller than that of malicious tampering. Because wavelet coefficients at each scale represent components having different frequencies, the amounts of modification on wavelet coefficients at each scale are different based on the type of incidental modification. By integrating the tamper response at each scale, we propose a unified approach to distinguish malicious tampering from incidental distortion.

2. MEAN QUANTIZATION

In watermarking, quantization-based approach is the simplest one because it requires the least storage. In addition, it is oblivious by nature. However, a conventional quantization-based approach is very sensitive to image modification and cannot distinguish incidental distortion from malicious tampering. Usually, a quantization-based blind watermarking approach [4] divides a real number axis into multiple uniform intervals, and then assigns wa-

0-7803-6297-7/00/$10.00

termark symbol to each interval periodically. Given a quantization level q, a real value x can be quantized as $x = \lfloor \frac{x}{q} \rfloor \cdot q + r$, where q is a quantization level and $0 \leq r < q$ is a quantization noise. (To preserve the visual quality of watermarked image, the modifications of wavelet coefficients should not exceed the marking threshold [9].) Assume there are two watermark symbols, the quantization function is defined as: $Q(x,q) = S_k$ if $\lfloor \frac{x}{q} \rfloor \mod 2 = k$, where x is a real value, q is a quantization level and $S_k, k = 0, 1$ are the watermark symbols. For binary watermark, the embedding rules are as follows. In case of target watermark 1, if $Q(x,q) = 1$, x is unchanged. If $Q(x,q) = 0$, then x is increased or decreased by q, such that the new value x' satisfies $Q(x',q) = 1$. Similarly, in case of target watermark 0, a similar update rule can be applied. Kundur and Hatzinakos [4] have used a quantization approach in the wavelet transform domain to perform image authentication.

The reason why they designed the new mechanism is based on the assumption that any modification on image will lead to the change of corresponding wavelet coefficients and watermarks. By examining the extracted watermark, the areas with watermark errors are marked as tampered areas.

Let Δ^I and Δ^M denote the amount of tampering caused by incidental distortion and malicious tampering, respectively. In the case of incidental modification, Δ^I can be modeled as a Gaussian distribution with small variance, that is $\Delta^I \sim N(0, \sigma^I)$, where σ^I denotes the variance of an incidental distortion. On the other hand, in malicious tampering, Δ^M can be modeled as a Gaussian distribution with large variance, that is $\Delta^M \sim N(0, \sigma^M)$, where σ^M denotes the variance of a malicious tampering. Here, we have made an important assumption, i.e. $\sigma^I < \sigma^M$ according to [5]. Figure 1 illustrates the Gaussian distributions of tampering on wavelet coefficients due to incidental modification and malicious tampering. The probability of watermark error is $P(|\Delta| > 0.5 \times q)$. As shown in Figure 1, the probability of watermark error due to incidental modification is smaller than malicious tampering because incidental modification produces comparatively smaller variance. If the variance of incidental modification can be further reduced, the probability of watermark error can be reduced.

Given a set of samples, the population mean has smaller variance than that of individual samples. The proposed mean quantization approach embeds a watermark by revising the mean of weighted wavelet coefficients such that the variation is relatively small even if the changes on individual samples are large. In incidental distortion, the magnitudes of wavelet coefficients are more important than their sign. Assume $\{x_i\}$ is a set of wavelet coefficients, a *weighted mean* is defined as

$$\hat{x} = \sum_{i=1}^{n} (-1)^i |x_i|, \tag{1}$$

where $i = 1, \ldots, n$ and n is the number of coefficients. In Equation (1), the sign of each x_i is discarded, and each coefficient is weighted by artificial sign, $(-1)^i$. This arrangement has the advantage of preserving small variation when incidental distortions such as equalization, blurring, sharpening, and JPEG compression are encountered. For example, if the magnitudes of all x_i's are increased by Δ due to some high-pass processing, then

$$\begin{aligned} \hat{x}' &= \sum_{i=1}^{n} (-1)^i (|x_i| + \Delta) = \sum_{i=1}^{n} (-1)^i |x_i| + \sum_{i=1}^{n} (-1)^i \Delta \\ &= \sum_{i=1}^{n} (-1)^i |x_i| = \hat{x}, \end{aligned}$$

where n is an even number. Similar outcome holds for the case of low-pass processing.

In this paper, the watermark embedding process is based on the quantization function Q. The value $\hat{x}$ remains unchanged if the target watermark symbol is the same as $Q(\hat{x}, q)$. Otherwise, $\hat{x}$ should be increased or decreased with an amount q, such that the new watermark symbol, $Q(\hat{x}', q)$, is the same as the target watermark symbol. To update weighted mean $\hat{x}$, each individual coefficient x_i must be updated accordingly. The amount of update on x_i depends on its sign and weight. For example, to modify the magnitude $|\hat{x}|$ by Δ, the following rule can be applied to update $x_i, i = 1, \ldots, n$,

$$x_i' = x_i + (-1)^i \times sign(x_i) \times \Delta, \tag{2}$$

where $sign(x) = 1$ if $x \geq 0$, and $sign(x) = 0$ if $x < 0$. If the sign is changed after applying equation (2), i.e., $sign(x_i') \neq sign(x_i)$, then x_i' is set to 0. The reason of this arrangement is that the minimum value of magnitude cannot be smaller than zero.

For the amount of tampering, the distribution of magnitudes, $|\Delta|$, is different from its value, Δ. Let $x_i' = x_i + \Delta_i$ be a tampered coefficient, where x_i is the original coefficient and Δ_i is the amount of tampering. Because $\Delta_i \sim N(0, \sigma)$, the probability distribution function can be formulated as

$$p_N(\Delta_i) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{\Delta_i}{\sigma})^2}.$$

Let Y_i be the amount of change on $|x_i|$, i.e. $|x_i'| = |x_i| + Y_i$, and $-|x_i| \leq Y_i \leq \infty$. Because $x_i' = x_i + \Delta_i$ and $|x_i'| \geq 0$, the probability distribution of Y_i is

$$\begin{aligned} p(Y_i) &= p_N(\Delta_i) + p_N(2|x_i| - \Delta_i) \\ &= \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{\Delta_i}{\sigma})^2} + \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{2|x_i| - \Delta_i}{\sigma})^2}. \end{aligned}$$

Under the circumstances, the expected value μ_i of Y_i is greater than 0, and the variance σ^y is smaller than σ. That is, $(\sigma^y)^2 = E(Y_i^2) - \mu_i^2 \leq E(\Delta_i^2) = (\sigma)^2$. For weighted mean, the distribution becomes $N(0, \frac{1}{n}\sigma^y)$, where n is the number of coefficients used to embed a watermark. The advantage of weighted mean-based quantization is that when an incidental distortion decreases or increases wavelet coefficients uniformly, the weighted mean will not be changed significantly. However, there is a tradeoff between robustness and resolution.

3. TAMPERED AREA ESTIMATION USING INFORMATION FUSION

A possible way to estimate the maliciously tampered area is to integrate the detection results from all scales. When malicious tampering is applied, the amount of changes on wavelet coefficients are large at both coarse and fine scales. If the amount of modification on wavelet coefficient exceeds a quantization level,

this coefficient is treated as a tampered coefficient. However, the quantization-based approach assigns a same watermark symbol to multiple intervals, the coefficients having significant modifications may not reflect the changes of watermark symbols. Thus, it is expected that the probability of watermark error caused by malicious tampering is $\frac{1}{2}$. On the other hand, the probability of watermark error caused by incidental distortion ranges from 0 to $\frac{1}{2}$. The accuracy of estimated tampered area at each scale is different. To compute a reliable and accurate shape of tampered area, the responses at each scale are weighted and integrated.

For the sake of estimating the distribution of tamper response, the minimum Chess-Board distance [10] between a tampered point x and its nearest tampered point y, $y \neq x$, is defined as *density*, $Density(x)$, of the coefficient x. If x is not tampered, then $Density(x) = 0$. Those tampered points with $Density(\cdot) = 1$, are called *dense* and other tampered points with $Density(\cdot) > 1$, are called *sparse*. Dense points are defined to be malicious, so a rule of extracting the malicious tampered area is to group those dense points. Those tampered points are called point tamper responses. In addition, we define N_l^{total}, N_l^{tamper}, N_l^{dense}, and N_l^{sparse} as the total number of coefficients, the total number of tampered coefficients, the number of dense tampered coefficients and the number of sparse tampered coefficients, respectively, at each scale l. The value of N_l^{tamper} can be computed by counting the number of tampered coefficients with density greater than zero. Those tampered coefficients can be further classified into dense and sparse set. So, the relation, $T_l^{tamper} = T_l^{dense} + T_l^{sparse}$, holds.

In this paper, an estimation of the tampered area is the tampering ratio (TR), which is defined as $TR_l = (2.0 \times N_l^{dense})/N_l^{total}$. Because the probability of watermark error in malicious tampered area is about $\frac{1}{2}$, it is estimated that the number of malicious tampered coefficients would be twice of N_l^{dense}. Another measure of the importance at scale l is $WGT_l = N_l^{dense}/(N_l^{tamper})^2$. The definition of WGT_l is based on the observation that smaller tamper response, N_l^{tamper}, implies smaller N_l^{sparse}.

To emphasize the importance of dense points and suppress the sparse response, the point tampering responses at each scale are transformed into tamper response map (TRM) based on their Chess-Board distances among tampered coefficients. The tamper response function (TRF) of a tampered point $x_l(i^*, j^*)$ is defined as

$$TRF(x_l(i^*, j^*), x_l(i, j)) = \frac{\max\{|i^* - i|, |j^* - j|\}}{(\sum_{k=1}^{(Density(x_l(i^*, j^*)+1)} k^2)},$$

if $\max\{|i^* - i|, |j^* - j|\} \leq (Density(x_l(i^*, j^*))+1)$. If $x_l(i^*, j^*)$ is not tampered or the Chess-Board distance between $x_l(i^*, j^*)$ and $x_l(i, j)$ is greater than $Density(x_l(i^*, j^*))$, then $TRF(x_l(i^*, j^*), x_l(i, j))$ is equal to 0. Each tampered coefficient has its corresponding tamper response. All tamper responses are integrated to form the tamper response map, i.e., $TRM_l(i, j) = \sum_{i^*, j^*} TRF(x_l(i^*, j^*), x_l(i, j))$. The tamper response maps at each scale are weighted by WGT_l and integrated to form the final tamper response map, i.e. $TRM^{final}(i, j) = \sum_l WGT_l \times TRM_l(i, j)$. Let l^* denote the scale with minimum TAF, then the ratio of tampered area over entire image is TR_{l^*}. If we sort the values of $TRM^{final}(i, j)$ in a decreasing order, then the tampered areas are indicated by those points with large value of $TRM^{final}(i, j)$.

During the information fusion of multiscale point tamper response, the following rules can be applied to distinguish malicious tampering and incidental distortion.

- **Rule 1:** If $TAF_l = 0$ at all scale l, then the watermarked image is neither maliciously tampered nor incidentally distorted.
- **Rule 2:** If $TAF_l = 0$ at some scale l, then the watermarked image has been processed by only some incidental distortions.
- **Rule 3:** Let s^* denotes the scale such that $TAF_{l^*} = \min_l\{TAF_l\}$. If $TAF_{l^*} > 0$, and $N_{l^*}^{dense} < \alpha \times N_{l^*}^{tamper}$, where the range of constant α is $0.5 \leq \alpha \leq 1.0$, then the watermarked image has been processed by only incidental distortion.
- **Rule 4:** If $N_l^{dense} = N_l^{tamper}$, at all scale l, then the watermarked image is only maliciously tampered.
- **Rule 5:** If all the above rules do not hold, then the watermarked image is both maliciously tampered and incidentally distorted.

The success of the above decision rules relies on the relative degree of point tamper response between malicious tampering and incidental distortion. However, it is difficult to distinguish malicious tampering and incidental distortion when the following conditions occurred. If the tampered area is too small or many small tampered areas are very small, the decision rules will fail to identify the source of tampering. To detect this type of malicious tampering, the user should examine those point tampering response at each scale manually. Only if the probability of watermark error caused by incidental distortion is zero, we can claim that the small area tamper response generated by malicious tampering. The number of scale and the number of coefficients used for watermarking are two important factors, which will affect the entire performance. When the number of scale increases, the resolution decreases and the allowable modification of wavelet coefficients decreases.

4. EXPERIMENTAL RESULTS

The proposed method is tested using the Pepper image of size 512×512, as shown in Figure 2(a). In the experiment, 4 scale wavelet transform is performed and each mean coefficient corresponds to 16 wavelet coefficients. The watermarked image is shown in Figure 2(b) with PSNR 35.91 dB. The image with two artificial objects and the superimposed image are shown in Figure 2(c) and (d), respectively. The tampering image will be used to simulate the malicious tampering in the following experiments.

Figure 3 illustrates the intermediate results of a maliciously tampered image verified using the proposed method. Figure 3(a) is the image maliciously tampered and incidentally distorted by JPEG compression with quality factor 90. Those areas corresponding to watermark errors at scales 1 to 4 are depicted in Figure 3(b)-(e). At scale 1 to scale 4, the area corresponding to malicious tampering all have dense point tamper response. Those point tampered responses are then transformed into tamper response maps as shown in Figure 3(g)-(j). We can see that sparse response corresponding to incidental distortion will have weak response as shown in Figure 3(f). These tamper response maps are weighted by WGT^s and integrated to form final tamper response map, shown in Figure 3(f). In this example, the superimposed object is correctly located.

Further experimental results about maliciously tampered and incidentally distorted images are as follows. First, the tampered image is blurred using a 3×3 mask. The detection result is shown in Figure 4(a). Because the behavior of blurring is similar to a mean

quantization, the tampered area is correctly identified. Second, the watermarked image is not only tampered but also equalized. Because the amounts of modification on wavelet coefficients are pretty large, most areas of image are treated as maliciously tampered. Third, the tampered image is further JPEG compressed with quality factor 50 and the detection result is shown in Figure 4(c). Due to the behavior of JPEG compression tends to decrease the magnitudes independently, the proposed approach can identify the tampered area correctly. Finally, the tampered image is sharpened with factor 0.5. The detection result is shown in Figure 4(d). Once again, the tampered area is correctly identified.

From the above experiments, it is verified that the assumption of small variance and large variance with respect to malicious modification and incidental distortion, respectively, is reasonable. Using mean quantization-based watermarking, the detected tampering regions resulted from incidental distortion are quite small, but malicious tampering is detected at each level. In addition, when the number of coefficients used in mean quantization is increased, the robustness of the watermark is increased, but at the expense of little fragility.

5. CONCLUSION

In this paper, a mean quantization blind watermarking approach has been presented. The proposed method is able to perform image authentication even when images are JPEG compressed and then maliciously tampered. Statistical analysis and experimental results have proven that the proposed method is indeed superb. Future work will focus on analyzing the effect of weighted mean quantization on the tradeoff between robustness and fragility.

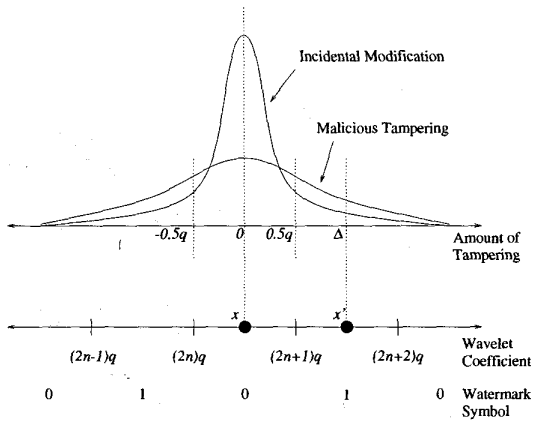

Fig. 1. The statistical distribution of incidental modification and malicious tampering on wavelet coefficients.

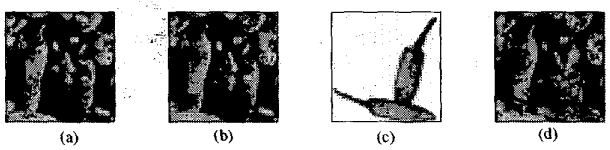

Fig. 2. The process of malicious tampering

Fig. 3. Example of tampered area response.

Fig. 4. Some detection results

6. REFERENCES

[1] S. Bhattacharjee and M. Kutter, "Compression tolerant image authentication," in *IEEE International Conf. on Image Processing*, Chicago, USA, October 1998.

[2] J. Dittmann, A. Steinmetz, and R. Steinmetz, "Content-based digital signature for motion pictures authentication and content-fragile watermarking," in *IEEE Inter. Conf. Multimedia Computing and Systems*, Italy, 1999, vol. II.

[3] G. L. Friedman, "The trustworthy digital camera: Restoring credibility to the photographic image," *IEEE Trans. on Consumer Electronics*, 1993.

[4] Deepa Kundur and Dimitrios Hatzinakos, "Digital watermarking for telltale tamper proofing and authentication," *Proceedings of the IEEE*, vol. 87, no. 7, pp. 1167–1180, July 1999.

[5] C.-Y. Lin and S.-F. Chang, "A robust image authentication method surving jpeg lossy compression," in *SPIE International Conf. on Storage and Retrieval of Image/Video Database*, San Jose, USA, January 1998, vol. 3312, EI'98.

[6] C. S. Lu, H. Y. Mark Liao, and L. H. Chen, "Multipurpose audio watermarking," *to appear in 15th Int. Conf. on Pattern Recognition*, Spain, 2000.

[7] C. S. Lu, H. Y. Mark Liao, and C. J. Sze, "Combined watermarking for image authentication and protection," *to appear in Proc. 1st IEEE Int. Conf. on Multimedia and Expo*, USA, 2000.

[8] C. S. Lu, H. Y. Mark Liao, S. K. Huang, and C. J. Sze, "Cocktail watermarking on images," in *Proc. 3rd Inter. Workshop on Information Hiding*, Dresden, Germany, 1999, LNCS 1768, pp. 333–347.

[9] Andrew B. Watson, Gloria Y. Yang, Joshua A. Solomon, and John Villasenor, "Visibility of wavelet quantization noise," *IEEE Trans. on Image Processing*, vol. 6, no. 8, pp. 1164–1175, August 1997.

[10] Ramesh Jain, Rangachar Kasturi, and Brian G. Schunck, *Machine Vision*, MacGraw-Hill, Inc., 1995.