INFORMED AUTHENTICATION WATERMARKING VIA STEGO DATA RECONSTRUCTION

Chao-Yong Hsu[1,2] *and Chun-Shien Lu*[2,*]

[1]Graduate Institute of Communication Engineering, National Taiwan University, Taipei, Taiwan, ROC
[2]Institute of Information Science, Academia Sinica, Taipei, Taiwan, ROC

ABSTRACT

Media authentication aims to judge the integrity of media content in the sense that malicious tampering should be detected while incidental modifications should be tolerated. This study investigates the resistance of a semi-fragile watermarking method to incidental manipulations with focus on resisting compressions with lower bit rates. To this end, we develop an informed authentication watermarking scheme based on reconstructing transformed-domain data in the sense that the effect resulted from incidental modifications can be eliminated through reconstruction. Statistical analyses and experimental results are provided to validate the proposed method.

1. INTRODUCTION

Watermark-based image authentication approaches are helpful to ensure the credibility of digital media by detecting potential tampering based on the (semi-)fragility of hidden watermarks. To effectively satisfy both objectives of robustness and fragility, Kundur *et al.* [1, 2] designed a new framework for digital watermarking, which is semi-fragile to any form of acceptable distortions, random or deterministic, such that both objectives of robustness and fragility can be effectively controlled and achieved. However, the capability of resistance to JPEG compressions is limited to larger quality factors (e.g., $\geq 70\%$). Lin and Chang [3, 4] dealt with the problems of error detection and recovery of still images. They proposed a pre-quantization technique to adjust the DCT coefficients used for embedding in advance. Then, based on a desirable robustness against compressions, the DCT coefficients are quantized again to accomplish embedding. Although this pre-quantization scheme can guarantee to resist compressions to a pre-defined level, the trade-off between fidelity and robustness often restricts the achievable performance.

The objective here is to enhance the resistance of a semi-fragile authentication watermarking scheme to compressions while maintaining fidelity. We examine the characteristics of DCT coefficient distributions under compression-free and compression-imposed environments, respectively, and exploit them to design a new paradigm of semi-fragile authentication watermarking. Since the prior knowledge about the distribution of transformed coefficients is exploited and reconstructed for watermarking, we, thus, call the proposed scheme "Informed Authentication Watermarking with Reconstruction (IAWR)." We provide statistical analyses and experimental results to validate IAWR.

This work was partially supported by NDAP-R&DTD-Digital Archives System Related Technology Research & Development Projcet: NSC 94-2422-H-001-007.*Contact Author (lcs@iis.sinica.edu.tw)

2. PROBLEM STATEMENT

In order to resist incidental modifications, the caused effects must be eliminated in order to recover the hidden watermarks. Traditionally, the distribution of high-frequency wavelet coefficients and non-DC DCT coefficients for an image is modeled as a Laplacian distribution with parameter λ, which is $f^L(x) = \frac{\lambda}{2}e^{-\lambda|x|}$. However, it can be observed that the distribution of DCT coefficients still deviates from the true Laplacian distribution up to a certain degree. In view of this, we propose to embed watermarks to enforce the approximate Laplacian distribution to become the true one. In other words, the pdf of DCT coefficients in a stego image can be equal to a true Laplacian distribution. When incidental manipulations (e.g., compressions) are encountered, the pdf of DCT coefficients in an attacked image is deviated from the true Laplacian distribution. Now, what we need to do is to reconstruct the deviated pdf back to the true form, from which the hidden watermarks can be successfully extracted.

3. STEGO DATA RECONSTRUCTION-BASED INFORMED AUTHENTICATION WATERMARKING

3.1. Invariant Sum of Magnitudes in a set of DCT Coefficients Before and After Incidental Manipulations

An image is divided into 8×8 blocks, denoted as $\mathbf{B_b}$. Each block $\mathbf{B_b}$ is DCT transformed to get DCT coefficients, $B_b(i,j)$ $(1 \leq i,j \leq 8)$. After compressions, let $B_b^q(i,j)$ denote the quantized coefficient and let $r_b(i,j) = B_b(i,j) - B_b^q(i,j)$ be quantization error. We study the relationship between $E(\sum_b \sum_{i,j} B_b^q(i,j))$'s and $E(\sum_b \sum_{i,j} B_b(i,j))$'s from another viewpoint; i.e., we consider the relationship between absolute transformed coefficients. Let the quantization error be expressed as $r_b(i,j) = |B_b(i,j)| - |B_b^q(i,j)|$. We can derive:

$$E(\sum_b \sum_{i,j} |B_b^q(i,j)|) = E(\sum_b \sum_{i,j} |B_b(i,j)|) - E(\sum_b \sum_{i,j} r_b(i,j)).$$

In order to eliminate the quantization errors, we propose a strategy, called reconstruction after quantization, to reconstruct the distribution of $B_b(i,j)$'s from those $B_b^q(i,j)$'s. The reconstructed DCT coefficients without quantization effect are denoted as $B_b^r(i,j)$'s. Based on the reconstructed distribution of $B_b^r(i,j)$'s, which is expected to be approximate that of $B_b(i,j)$'s, it can be further derived to obtain

$$E(\sum_b \sum_{i,j} |B_b(i,j)|) \approx E(\sum_b \sum_{i,j} |B_b^r(i,j)|), \quad (1)$$

which specifies the desired property: invariant sum of magnitudes in a set of DCT coefficients before and after incidental manipulations (e.g., compressions). Thus, the hidden watermark signals can be expected to remain without being destroyed. That is the key point in the proposed reconstruction-based semi-fragile authentication watermarking method.

The proposed scheme for piecewise reconstruction of $B_b(i, j)$'s distribution from $B_b^q(i,j)$'s is described in **Appendix A**. We will also prove how the reconstructed distribution can preserve invariant magnitude sum of DCT coefficients, which can be exploited to resist incidental modifications.

3.2. Transferring an Approximate Laplacian Distribution to a True One via Embedding

From the quantized coefficients, $B_b^q(i,j)$'s, their probability density function can be approximated using Laplacian distribution, and the corresponding parameters can be yielded. Our empirical observations find that both distributions of transformed coefficients before and after quantization preserve similar shapes so that the parameters of Laplacian distribution can be approximately obtained by modeling quantized coefficients. With the parameters, the theoretic probability density function can be derived to facilitate piecewise reconstruction of DCT coefficients' distribution.

In this study, all 8×8 DCT blocks in an image form a group of 64 subbands. Only the middle-frequency subbands are used for embedding in order to maintain fidelity. In addition, in order to achieve high-resolution authentication, we propose a strategy of interleaving embedding. For an image of size $Dim_X \times Dim_Y$, watermarks are embedded into horizontal slices of size $8 \times Dim_Y$ and vertical slices of size $Dim_X \times 8$. According to interleaving watermarking, the overlapping block is of size 8×8, which constitutes the so-called minimum authentication block.

For each slice, it is embedded with a watermark signal of 16 bits ; i.e., 16 middle-frequency subbands are selected for embedding and one subband (called a set hereafter) is inserted with one bit. Let Δ^q be the size of a quantization interval, which is used to quantize the "mean" of a set of DCT coefficients, and the quantization intervals are labeled with either 0 or 1 successively. For horizontal slice watermarking, the size of a set of DCT coefficients is $\lfloor \frac{Dim_Y}{8} \rfloor$, and is $\lfloor \frac{Dim_X}{8} \rfloor$ for vertical slice watermarking. It is important to note that the selected subbands used for embedding in the horizontal and vertical slices are different to avoid overlapped embedding that will incur interference of detection. With quantization-based watermarking, the mean of a set of DCT coefficients is moved to the middle of a quantization interval depending on the incoming watermark bit.

So far, the entire embedding process at this stage has not been carried out completely because watermark embedding only proceeds to the "group" level. In practice, each coefficient in a set still needs to be modified accordingly to satisfy the result generated from group-wise embedding, as described in the above. According to **Appendices A** and **B**, a Laplacian distribution parameter can be estimated for each quantization interval Δ^q. In practice, the pdf is continuous, but in order to facilitate implementation it is regarded to be discrete. Therefore, to accomplish coefficient-wise embedding with accurate piecewise approximation, the size of quantized intervals, Δ^c, for re-calculating Laplacian pdf is set to 1. Based on this setting, the DCT coefficients in an embedding interval Δ^q are properly modified step by step in terms of Δ^c so that the distribution of modified coefficients can further approximate the true Laplacian distribution. It can be said that the introduced distortions during embedding is produced by a perturbation procedure in that a noise is added into a subband's pdf to further approximate $f^L(\cdot)$.

3.3. Transferring Quantized Coefficients to Theoretic Laplacian Distribution for Extraction

For each slice of a suspect image, its distributions of 8×8 DCT coefficients for the selected 16 subbands are obtained. For each distribution, it is reconstructed back to the true Laplacian distribution to obtain a reconstructed version. From this, the label that specifies the interval into which each set of DCT coefficients falls is extracted to be a watermark bit. Therefore, the bit error rate (BER) corresponding to each slice can be generated. Let $BER_h(x)$ and $BER_v(y)$ denote the BER with respect to the x-th horizontal and the y-th vertical slice, respectively. So far, the integrity of a slice can be checked. However, as described previously, we propose an interleaving watermarking scheme in order to achieve high-resolution authentication. As a consequence, the BERs obtained from each pair of horizontal and vertical slices are combined to get an indicator for authentication of a block of size 8×8.

To check the integrity of an authentication unit of size 8×8, a simple measurement is defined to represent the integrated bit error rate, $BER_{x,y}$, of the authentication unit, which is the overlapped part resulted from a pair of the x-th horizontal slice and the y-th vertical slice: $BER_{x,y} = (BER_h(x) + BER_h(y)) \times 0.5$. While $BER_{x,y}$ may vary differently for the imposed operations, it nevertheless serves an indication as to whether a block is incidentally or maliciously manipulated. In this paper, if $BER_{x,y}$ is larger than a threshold ϵ, then the

block is judged to has been maliciously tampered with.

3.4. Analysis of Detection Performance

The performance of our piecewise reconstruction refers to resistance to incidental manipulations and fragility to malicious manipulations. We shall discuss the reconstruction error and its impact on performance. According to **Appendix A**, the reconstruction error e for a subband can be written as

$$e = |E(C) - E(C^r)| = \int_0^\infty c(|f(c) - f^r(c)|)dc = |\frac{\lambda - \lambda^r}{\lambda\lambda^r}|,$$

where λ^r is the parameter corresponding to the reconstructed Laplacian distribution.

It is said that a watermark bit hidden in a subband can be successfully extracted if e is smaller than $\frac{\Delta^q}{2}$; i.e.,

$$e = |\frac{\lambda - \lambda^r}{\lambda\lambda^r}| < \frac{\Delta^q}{2}. \quad (2)$$

We further let $\lambda^r = \tau\lambda$, where τ is a positive real number, and substitute it to Eq. (2). After some derivations, we have

$$\tau \in \begin{cases} (0 \ \frac{2}{2+\lambda\Delta^q}], & \text{if } 0 < \tau < 1, \\ (1 \ \frac{2}{2-\lambda\Delta^q}], & \text{if } \tau > 1. \end{cases} \quad (3)$$

As a result, we show that if the reconstructed λ^r satisfies the above constraints, then resistance to the imposed processing can be guaranteed. On the other hand, whenever the reconstructed λ^r with τ falling outside the ranges derived in Eq. (3), the imposed operations are judged to be malicious.

4. EXPERIMENTAL RESULTS

Two sets of experiments were conducted in this study. First, two cover images of size 512×512 were watermarked using our method and were compressed with various quality factors of JPEG, respectively. The PSNR values of the stego Lena and Baboon images were 40.11dB and 40.31dB, respectively. The BERs obtained from compressed images were shown in Table 1. We can observe that the BERs are significantly reduced when distribution reconstruction is employed to eliminate the effect caused by JPEG compressions. It is essential to note that the obtained BERs are nearly lower than 0.21% when QFs are as low as 30%.

Second, we verify the performance of proposed method in locating maliciously tampered regions under purely malicious modification and malicious + incidental modifications. Fig. 1(a) shows an image that has been watermarked using our method (PSNR is 37.95dB), and then its number plate "9C-9701" is removed and pasted with another plate "2A-9165" to simulate pure malicious tampering, as shown in Fig. 1(b). When our method was performed, the maliciously altered regions were located and represented with inverse gray-scale, as shown in Fig. 1(c). It is obvious to observe that the pasted number plate has been sufficiently located.

(a) (b) (c)

Fig. 1. **Authentication of malicious tampering: (a) stego image; (b) maliciously tampered image with the number plate replaced; (c) located tampered regions with inverse gray-scale.**

On the other hand, Fig. 1(a) is first JPEG compressed with QF 70% and 50%, respectively, and then the above malicious operation is conducted again. Fig. 2 shows the regions that are judged to have been maliciously tampered with. We can observe from Fig. 2(b) that some JPEG compressed areas are also detected. The reason we find is that the DCT distribution of smoother regions is easy to be significantly distorted by compressions. Basically, both the incidental and malicious manipulations are simultaneously imposed on a stego image, our method is still able to distinguish incidental modifications from malicious modifications. It is essential to note that to our knowledge this challenging problem has not been properly addressed in the literature, in particular when JPEG compression is with quality factor as low as 50%.

(a) (b)

Fig. 2. **Authentication of incidental+malicious manipulations: (a) detection of maliciously tampered regions from the image that has been JPEG compressed with QF 70% and replaced with another number plate; (b) similar to (a) but with QF 50%.**

5. CONCLUSIONS

The existing semi-fragile watermarking methods fail to resist compressions with larger compression ratios (or smaller quality factors in JPEG). In this paper, an informed authentication watermarking scheme is proposed based on reconstruction of transformed-domain information so that the effect resulted from incidental modifications (e.g., compressions) can be eliminated while that resulted from malicious distortions can be detected. Performance analyses and experimental results are provided to validate the proposed method.

Appendix A: *Piecewise Reconstruction of DCT Coefficients' Distribution and Its Invariance to Compressions*

Table 1. **Resistance of Proposed Informed Authentication Watermarking to JPEG. "w" and "w/o" denote watermark detection using and without using distribution reconstruction, respectively. QF, ranging from $90\% \sim 10\%$, means quality factor in JPEG.**

Lena	QF=90%	QF=80%	QF=70%	QF=60%	QF=50%	QF=40%	QF=30%	QF=20%	QF=10%
PSNR (dB)	36.31	34.17	33.53	33.09	32.75	32.34	31.83	30.93	29.06
BER (w/o)	0.000	0.010	0.058	0.298	0.491	0.628	0.700	0.534	0.544
BER (w)	0.000	0.000	0.046	0.101	0.110	0.151	0.213	0.342	0.443
Baboon	QF=90%	QF=80%	QF=70%	QF=60%	QF=50%	QF=40%	QF=30%	QF=20%	QF=10%
PSNR (dB)	34.52	31.12	29.23	28.00	27.14	26.35	25.46	24.33	22.61
BER (w/o)	0.000	0.000	0.024	0.166	0.311	0.502	0.648	0.556	0.492
BER (w)	0.000	0.000	0.012	0.090	0.053	0.098	0.217	0.357	0.456

In Fig. 3, it shows two DCT distributions where subfigure (a) shows the continuous stego DCT distribution and subfigure (b) shows the discrete quantized DCT distribution. For Fig. 3(a), let C be a random variable that theoretically takes on uncountable number of stego DCT coefficients and let its probability density function be denoted as $f(c)$, where the values of c are positive because we consider absolute values of DCT coefficients. Similarly, let C^q be a random variable representing the quantized but discrete DCT coefficients and its probability mass function is denoted as $p(c^q)$ for Fig. 3(b). The mean of magnitudes of the stego DCT coefficients can be expressed as $E(C) = \int_0^\infty cf(c)dc$.

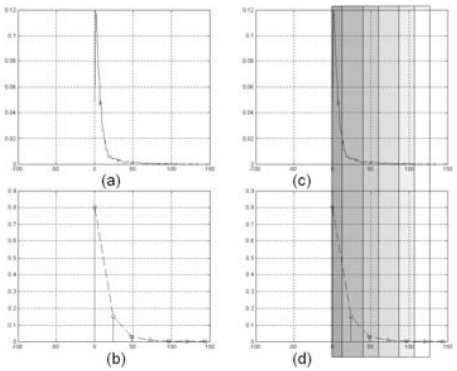

Fig. 3. **Distributions of absolute values of DCT coefficients: (a) continuous distribution before compression; (b) discrete distribution after compression; each interval of different gray-scale values in (c) and (d) denotes the basic unit used for piecewise reconstruction of (d) back to continuous distribution expected to similar to a true Laplacian pdf.**

In order to achieve piecewise reconstruction, we take one discrete value of Fig. 3(b) each time and from that a piece of continuous function is reconstructed as follows. By comparing the k-th quantization interval (colored in gray-scales) $[\frac{(2k-1)\Delta^q}{2} \quad \frac{(2k+1)\Delta^q}{2}]$ in Figs. 3(c) and 3(d), we have

$$c_k^q = \int_{\frac{(2k-1)\Delta^q}{2}}^{\frac{(2k+1)\Delta^q}{2}} f(c)dc. \tag{4}$$

From each c_k^q, a piece of continuous function, $f_k^r(c)$, can be constructed. In our study, for the size Δ^q of a quantization interval, it is further divided into finer interval of size Δ^c, which is set to 1 here. By doing so, a smoother distribution function can be yielded by piecewise composition.

It is important to note that since the information about Laplacian distribution and c_k^q are available, each constructed piecewise function $f_k^r(c)$ can be gradually generated in a discrete manner to fit the target, i.e., the Laplacian distribution. Thus, the function $f^r()$ can be generated as $f_0^r() \circ f_1^r() \circ \cdots \circ f_\infty^r()$. The most crucial task is to estimate the parameter, λ, in a Laplacian distribution (see **Appendix B**).

Once piecewise reconstruction is done, the mean of the reconstructed DCT coefficients can be expressed as $E(C^r) = \int_0^\infty cf^r(c)dc$, where $f^r(c)$ denotes the reconstructed probability density function. By means of the proposed piecewise reconstruction, we expect that the reconstructed mean, $E(C^r)$, is close to $E(C)$.

Appendix B: *Laplacian Distribution Parameter*

When c_k^q and the Laplacian distribution function are available, they can be related as

$$c_k^q = \int_{\frac{(2k-1)\Delta^q}{2}}^{\frac{(2k+1)\Delta^q}{2}} f^L(x)dx = F^L(\frac{(2k+1)\Delta^q}{2}) - F^L(\frac{(2k-1)\Delta^q}{2}),$$

where $F^L(x) = 1 - exp(-\lambda x)$ is a cumulative Laplacian density function. After some derivations, the value of λ_k corresponding to k-th interval can be calculated to be

$$\lambda_k = \frac{-2ln(1-(t_0+2(t_1+t_2+...+t_k)))}{(k+1)\Delta^q},$$

where $t_k = F(\frac{(2k+1)\Delta^q}{2}) - F(\frac{(2k-1)\Delta^q}{2})$.

6. REFERENCES

[1] C. Fei, D. Kundur, and R. Kwong, "Analysis and Design of Authentication Watermarking," *Proc. SPIE Security and Watermarking of Multimedia Contents VI*, Vol. 5306, 2004.

[2] D. Kundur, Y. Zhao and P. Campisi, "A Steganographic Framework for Dual Authentication and Compression of High Resolution Imagery," *Proc. IEEE Int. Symposium on Circuits and Systems*, Vancouver, 2004.

[3] C. Y. Lin and S. F. Chang, "SARI: Self-Authentication-and-Recovery Image Watermarking System," *Proc. ACM Multimedia Conf.*, 2001.

[4] C. Y. Lin, D. Sow, and S. F. Chang, "Using Self-Authentication-and-Recovery Images for Error Concealment in Wireless Environments," *SPIE ITCom/OptiComm*, Denver, CO, Vol. 4518, 2001.