MCiS 2006
Conference in technical cooperation with the IEEE and IEEE Computer Society

Detecting malicious attacks using semi-fragile watermark based on visual model

Hyunho Kang[*], Brian Kurkoski[**], Kazuhiko Yamaguchi[**], and Kingo Kobayashi[**]
{[*]Graduate School of Information Systems, [**]Dept. of Inf. and Communications Eng.},
University of Electro-Communications, 1-5-1 Chofugaoka, Chofu-shi, Tokyo 182-8585, Japan
{kang, kurkoski, yama, kingo}@ice.uec.ac.jp

***Abstract*—Semi-fragile watermarking is robust to mild modifications but fragile to malicious attacks. To achieve this goal, we use a watermarking system which exploits the features of human visual system (HVS). We can classify the nature of the attack by analyzing the number of non-extractible blocks statistically under the proposed method. Experimental results show that the proposed method is robust to unintentional modifications while is fragile to malicious attacks, such as JPEG compression 60% and less (where 100% is maximum quality).**

1. Introduction

Alteration of documents can occur by non-malicious modification or can be intentional attacks. The so-called unintentional or innocent alterations typically arise from modifications such as bit errors during transmission and storage, or signal processing operations such as filtering, contrast enhancement, sharpening, and compression. Intentional or malicious alterations, on the other hand, are assumed to be due to an explicit forgery attempt by a pirate for the purpose of removing the document's watermark [1].

In recent years, many semi-fragile watermarking systems have been proposed. The technique in [2] is applied to 64-by-64 blocks in the spatial domain, and is especially sensitive to smoothing processing and is computationally one of the most complex algorithms. The technique in [3] is applied to 4-by-4 blocks in wavelet domain, and is weak against most signal processing operations. The technique in [4] is applied to a pair of 8-by-8 blocks in the DCT domain, and is very fragile against all signal processing attacks except for JPEG compression. The technique in [5] is applied to the whole image or 64-by-64 blocks units in the spatial domain, and it is the weakest of all algorithms against any signal processing manipulations. The technique in [6] is applied in spatial domain, and is considered to be the most robust algorithm against JPEG compression. The technique in [7] are applied to 8-by-8 blocks in the DCT domain, and it can withstand signal processing operations except for histogram equalization and smoothing. The technique in [8] are applied in groups of 8-by-8 blocks in the DCT domain, and it is very fragile with respect to signal processing operations. Nakai et al. [9] improved Kundur et al. [3] method, and Maeno et al. [10] improved Chang et al.[4] method. Lin et al. [11] is robust against almost all non-malicious signal processing operations except smoothing.

Although many semi-fragile algorithms have been proposed, they can not completely classify the nature of the modification. In this paper, we can classify the nature of the attacks and compared to the result of Lin et al. method [11], which is a representative semi-fragile watermarking technique.

The rest of the paper is organized into the following sections: Section 2 describes the proposed algorithm, including the human visual system, watermark construction, watermark embedding and watermark detection. Experimental results and conclusions are given in Section 3 and 4, respectively.

2. Semi-Fragile Watermarking

We suggest a watermark construction and embedding method using the just noticeable differences (JND) visual model that was proposed by Watson [12], and a corresponding detection method using wavelet transforms. Although JND has been used in image-adaptive watermarking, our method is distinct because the watermark construction itself used JND and embedding strength is image adaptive. Research on semi-fragile watermarking often does not address the importance of watermark construction itself.

2.1 Human Visual System

In the proposed paper, we make a watermark using a threshold unit which is often called "just-noticeable differences," or JND [12]. Originally the JND scheme has

been applied in the image compression field. Recently, the JND scheme has been introduced as an adaptive watermarking technique [13].

Let J_{ijk} be a threshold (JND), the values making method is described as follows:

$$J_{ijk} = e_{ijk} / m_{ijk} \quad (1)$$

where e_{ijk} is (i,j)-th quantization error in the k-th block is given by eq. (2) and m_{ijk} is (i,j)-th contrast masking in the k-th block is given by eq. (3).

Each image block of size {8,8} is transformed into its DCT, which we write c_{ijk} is (i,j)-th DCT frequency in the k-th block. Each block is then quantized by dividing it, coefficient by coefficient, by a quantization matrix q_{ij}. The quantization error e_{ijk} in the DCT domain is then

$$e_{ijk} = c_{ijk} - (\lfloor c_{ijk} / q_{ijk} + 0.5 \rfloor) q_{ij} \quad (2)$$

$$m_{ijk} = Max[t_{ijk}, |c_{ijk}|^{w_{ij}} \; t_{ijk}^{\,1-w_{ij}}] \quad (3)$$

where t_{ijk} is (i,j)-th luminance masking in the k-th block is given by eq. (4) and w_{ij} is a constant value equal to 0.7.

$$t_{ijk} = t_{ij} (c_{00k} / \hat{c}_{00})^{a_T} \quad (4)$$

where t_{ij} is (i,j)-th frequency sensitivity is given by $q_{ij}/2$, c_{00k} is the DC coefficient of the DCT for block k, $\hat{c}_{00}$ is the mean value of luminance (1024 for an 8 bit image), and a_T is a constant value equal to 0.649.

2.2 Watermark construction

First we construct a watermark, Wk, where 1≤ k ≤t and t is the number of total blocks (t=M/8*N/8, an original image of size $M \times N$), from a pseudo-random floating point sequence consisting of an array of 8-by-8 numbers, Gaussian distributed with average 0 and variance 1, R=$\{r_{ij}\}$, -3≤ r_{ij} ≤3, i=1,2,…,8 and j=1,2,…8. Fig.1 shows the creation process of the watermark. We find the JND threshold value of the discrete cosine transform (DCT) of R, C=DCT(R), using eq. (1). The portion of the DCT for which C is less than J=JND(C) is set to zero. The watermark is constructed from the inverse DCT (IDCT) of this.

2.3 Watermark embedding

In the proposed method, an 8-by-8 watermark is inserted in the spatial domain by adaptive strength α. This is especially noteworthy in the case of semi-fragile watermarking. We define X_k as the original image block, of size 8-by-8; define Y_k as the watermarked image block is given by eq. (5), of size 8-by-8; define W_k as the watermark from Fig. 1 and define α as the embedding strength:

$$\alpha = \begin{cases} 3 & if\ DCT(X_k) \geq JND(DCT(X_k)) \\ 0.5 & otherwise \end{cases}$$
$$Y_k = X_k + \alpha \cdot W_k \quad (5)$$

Fig. 2 shows the original image X, watermark W composed of W_k, and watermarked image Y. Observe that the watermark is distributed in all area of the image.

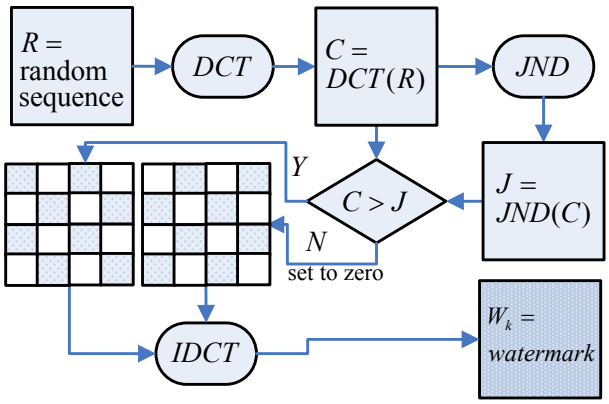
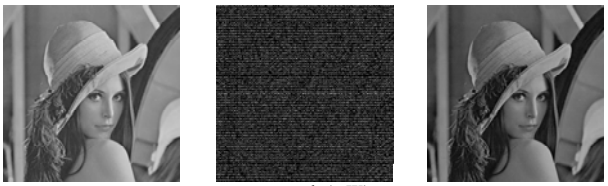


Fig. 1 Watermark construction.



original image (X) watermark (αW) watermarked image (Y)

Fig. 2 Original, watermark and watermarked image (PSNR: 42dB).

2.4 Watermark detection

The goal of this research is to detect the presence or absence of the watermark on a block-by-block basis as a way of measuring the effectiveness of the proposed algorithms. We use the wavelet transform and linear correlation to detect the presence or absence of the watermark on a block-by-block basis. Each block of the attacked image is divided into low and high frequency coefficients by the discrete wavelet transform (DWT). The low frequency portion is set to zero. The signal, with the low frequency coefficient set to zero is processed by inverse DWT (A_k). If the correlation value ($corr_k$) is less than some threshold T, then it is a non-detected block (I_k=1). The detector counts the number of non-detected blocks in the image, and this number (S) is used to estimate if the image modification was malicious or non-malicious.

$$corr_k = \frac{1}{64}\sum A_k \cdot W_k \tag{6}$$

$$S = \sum_{k=1}^{4096} I_k \tag{7}$$

Experimentally, we found T=0.001 to be best in the sense of classifying JPEG compression ≤ 60% as malicious, and 70% as non-malicious.

Fig. 3 shows the block diagram of the watermark detection and counting a non-detected block. In our experiment, if $S > 40$ in the attacked image, then regard it as malicious attack (see Table.1).

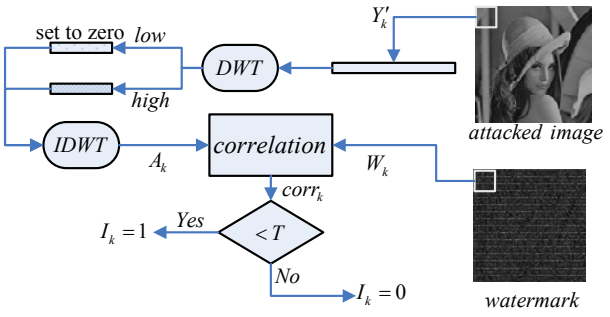


Fig. 3 Watermark detection (Let Y'_k be the attacked Y_k)

3. Experimental Result

In this paper, we consider the following list of manipulations as examples of non-malicious attacks [1]. We expect that our algorithm should find a low number of non-detected blocks in images with these alterations.

- Median filtering with a support of 3×3.
- Salt-and-pepper noise, up to one percent.
- Histogram equalization (uniform distribution).
- Sharpening (up-sharp masking filter with $\begin{bmatrix} -1 & -1 & -1 \\ -1 & 9 & -1 \\ -1 & -1 & -1 \end{bmatrix}$).
- Low-pass filtering 3×3 (equal weight coefficients equal to 1/9).
- Additive Gaussian noise with a signal-to-noise ratio of 35 dB.
- Mild compression, for example up to 70% JPEG.

In this paper, JPEG compression ≤60% is considered a malicious attack.

We considered the blocks with a non-detected watermark, in Fig. 4 and 5 the corresponding block is marked white; Fig. 6 plots the total number of such blocks for various attacks. With the Lin et al. method, it is difficult to determine the malicious attacks, because the image with non-malicious attacks has many non-detected blocks. However, our method clearly shows that the malicious attacks are those with JPEG compression 60% and less.

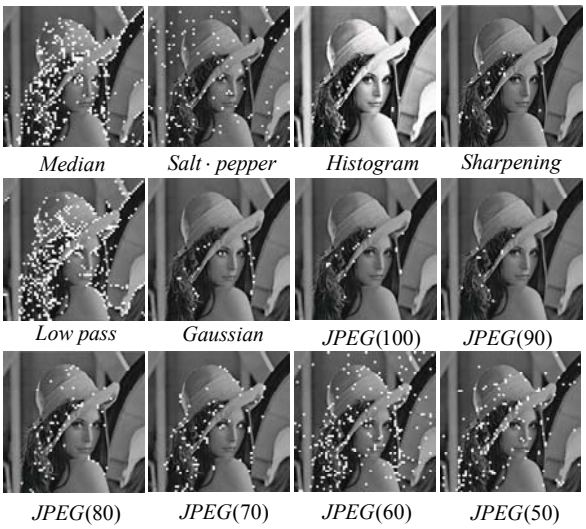


Fig. 4 Lin *et al*. method (the white blocks mean non-detected watermark blocks)

Fig. 5 Our method (the white blocks mean non-detected watermark blocks)

The objective is to get a low number of non-detected blocks in non-malicious attacks and a high numbers of non-detected blocks in malicious attacks; this is our method's result as in seen in Fig. 5.

TABLE 1. THE NUMBER OF WHITE BLOCKS (NON-DETECTED WATERMARK BLOCKS)

Attacks	Intention*	Lin *et al*. method		Our method	
		# of non-detected block	Detected Intention*	# of non-detected block	Detected Intention*
Median filtering	N	402	M	3	N
Salt & pepper noise	N	127	M	0	N
Histogram equalization	N	30	N	2	N
Sharpening	N	27	N	0	N
Low pass filtering	N	447	M	5	N
Gaussian noise	N	21	N	0	N
JPEG(100)	N	18	N	0	N
JPEG(90)	N	12	N	0	N
JPEG(80)	N	66	M	0	N
JPEG(70)	N	94	M	5	N
JPEG(60)	M	252	M	65	M
JPEG(50)	M	156	M	199	M

* The intention of attacks: malicious attack (M) and non-malicious attack (N)

Table 1 shows the number of blocks whose watermarks could not be extracted using Lin *et al*. and our method against 12 attacks.

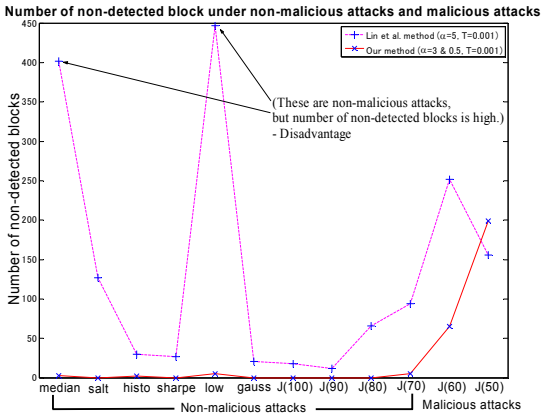


Fig. 6 Lin *et al*. method (+) vs. Our method (×)

4. CONCLUSION

As shown in the experimental results, we can determine the nature of attacks by counting the number of non-detected blocks. This paper found that our method is more robust to non-malicious attacks and more fragile to malicious attacks, compared to the Lin et al. method.

REFERENCES

[1] O. Ekici, B. Sankur, B. Coskun, U. Naci, M. Akcay, "Comparative evaluation of semifragile watermarking algorithms," Journal of Electronic Imaging, Vol.13. Jan. 2004, pp. 209-216.

[2] J. Fridrich, "Methods for Tamper Detection in Digital Images," Proc. of ACM Workshop on Multimedia and Security, Oct. 1999, pp. 19-23.

[3] D. Kundur and D. Hatzinakos, "Digital Watermarking for Telltale Tamper-Proofing and Authentication," Proc. of IEEE Special Issue on Identification and Protection of Multimedia Information, Vol. 87, July. 1999, pp. 1167-1180.

[4] C.Y. Lin and S.F. Chang, "Semi Fragile Watermarking for Authentication JPEG Visual Content," Proc. of SPIE Int. Conf. Security and Watermarking of Multimedia Contents 2, Vol. 3971, Jan. 2000.

[5] L.M. Marvel, G.W. Hartwig and Jr C. Boncelet, "Compression Compatible Fragile and Semi-Fragile Tamper Detection," Proc. of SPIE Int. Conf. Security and Watermarking of Multimedia Contents 2, Vol. 3971, Jan. 2000.

[6] M.P. Queluz and P. Lamy, "Spatial Watermark for Image Verification," Proc. of SPIE Int. Conf. Security and Watermarking of Multimedia Contents 2, Vol. 3971, Jan. 2000.

[7] J.J. Eggers and B. Girod, "Blind Watermarking Applied to Image Authentication," Proc. of ICASSP Int. Conf. On Acoustics, Speech and Signal Processing, May. 2001.

[8] T.H. Lan, M.F. Mansour and A.H. Tewfik, "Robust High Capacity Data Embedding," Proc. of ICASSP Int. Conf. On Acoustics, Speech and Signal Processing, May. 2001.

[9] Y. Nakai, "Multivalued Semi Fragile Watermarking," Proc. of SPIE Int. Conf. Security and Watermarking of Multimedia Contents 4, Vol. 4675, 2002.

[10] K. Maeno, Q. Sun, S.F. Chang and M. Suto, "New Semi-Fragile Image Authentication Watermarking Techniques Using Random Bias and Non-Uniform Quantization," Proc. of SPIE Int. Conf. Security and Watermarking of Multimedia Contents 4, Vol. 4675. 2002.

[11] E. T. Lin, C.I. Podilchuk and E.J. Delp, "Detection of Image Alterations Using Semi-Fragile Watermarks," Proc. of SPIE Int. Conf. Security and Watermarking of Multimedia Contents 2, Vol. 3971, Jan. 2000.

[12] A.B. Watson, "DCT quantization matrices visually optimized for individual images," Human Vision, Visual Processing, and Digital Display IV, Proc. of SPIE, 1913-1914. 1993

[13] C.I. Podilchuk and W. Zeng, "Image-Adaptive Watermarking Using Visual Models," IEEE J. Select. Areas Commun., Vol. 16. May. 1998, pp. 525-539.