Watermarking Scheme Capable of Resisting Sensitivity Attack

Xinpeng Zhang and Shuozhong Wang

***Abstract*—In sensitivity attacks, attempts across the boundary between "watermarked" and "not watermarked" using an available detector can provide adequate information about an embedded watermark for clearing it without causing serious distortion. This letter proposes a novel watermarking system capable of resisting such sensitivity attacks, which contains a tailor-made embedding algorithm and a corresponding detecting method. Although an attacker can still find the most effective way for destroying the watermark in a constructed signal near the boundary, it is impossible to move a watermarked product away from the "watermarked" region with low distortion. Therefore, the watermark detector can be widely distributed without fear of being used for sensitivity attacks.**

***Index Terms*—Detector, digital watermarking, sensitivity attack.**

I. Introduction

With the development of watermarking techniques for protection of intellectual property rights, hostile attacks are attempted to invalidate the watermarking systems [1], [2]. In a considered scenario where a watermark detector is a black box available to all users, a pirate can perform a successful *sensitivity attack* by exploiting the detector without any knowledge about its internal mechanism [3], [4]. The pirate constructs a multimedia signal that is close to the decision boundary between the "watermarked" and "not watermarked" regions, and a great number of attempts across the boundary using the detector are made to provide adequate information about the embedded watermark for the purpose of removal with very little distortion. The number of attempts required is in the order of $O(N)$, where N is the number of data samples in the watermarked signal. This attack is also termed *oracle attack* or *detector attack*, and some more effective iterative methods are presented in [5] and [6].

Several approaches to resisting the sensitivity attack have been proposed. The countermeasure described in [4] is based on randomization of the detected results in a defined area between "watermarked" and "not watermarked" regions, but it is not secure enough [6], [7]. It is also suggested to convert the decision boundary to a fractal curve [5]. However, the fractalization does not change the outline of the decision boundary; therefore, it is still possible to estimate the embedded watermark signal in order to destroy it. In [8], the detected result of a digital product is randomized or delayed when a similar product is input to the detector; thus, repeated attempts in sensitivity attack are invalidated. In this approach, however, the watermark detector must employ and update a hash table to memorize the detected products. This means the watermark detector can only be installed on a dedicated server but not duplicated on distributed sites.

This letter proposes a watermarking scheme capable of defeating the sensitivity attack, in which a novel tailor-made embedding algorithm and a corresponding detection mechanism are designed to "mislead" the attackers. When an attacker tries to estimate the embedded watermark, he will get a "fake" signal. Adding/subtracting the "fake" estimation to/from a watermarked product does not affect the embedded watermark but only provides a clue indicating that the product has been tampered by a sensitivity attacker.

Manuscript received April 19, 2006; revised July 16, 2006. This work was supported in part by the National Natural Science Foundation of China under Grants 60372090 and 60502039 and in part by the Key Project of Shanghai Municipality for Basic Research under Grant 04JC14037. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Min Wu.

The authors are with the School of Communication and Information Engineering, Shanghai University, Shanghai 200072, China (e-mail: xzhang@staff.shu.edu.cn; shuowang@staff.shu.edu.cn).

Digital Object Identifier 10.1109/LSP.2006.882092

II. Secure Watermarking Scheme

A. Sensitivity Attack

Assume that a pseudo-random sequence is added to a host media to produce a watermarked copy, and the output of a watermark detector is either "watermarked" or "not watermarked," determined by comparing correlation between the pseudo-random sequence and the multimedia data with a threshold. In fact, the decision boundary between the "watermarked" and "not watermarked" regions is a hyper-plane in a multidimensional space. If an attacker possesses a watermarked copy and an available detector, he can remove the embedded watermark in the following steps.

1) By modifying the watermarked copy, the attacker can create a signal very close to the decision boundary between the "watermarked" and "not watermarked" regions.
2) Assuming that the number of samples is N, the constructed signal can be considered an N element vector. The attacker finds N orthogonal vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_N$ of length N (in a special case, the N orthogonal vectors correspond to N different pixels) and adds them to or subtracts them from the constructed signal with increasing strengths $\alpha_n (n = 1, 2, \ldots, N)$ until the detector output is flipped. The parameter $1/\alpha_n$ shows how sensitive the detector is to modification in the direction of each vector, and another parameter β_n that is either $+1$ or -1 indicates addition or subtraction, respectively.
3) The attacker combines the obtained knowledge about sensitivities to get

$$\mathbf{v}_{\mathrm{A}} = \sum_{n=1}^{N} (\mathbf{v}_n \cdot \beta_n / \alpha_n) \tag{1}$$

and subtracts $\mathbf{v}_{\mathrm{A}}$ from the watermarked copy with an increasing strength α_{A} until the detector reports that no watermark is present. Thus, the embedded watermark is removed.

In fact, $\mathbf{v}_{\mathrm{A}}$ is perpendicular to the decision boundary. In other words, it is proportional to the gradient of the correlation function and the embedded pseudo-random sequence, therefore

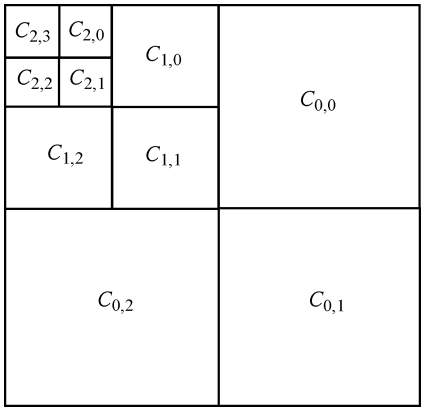

Fig. 1. Decomposition of DWT.

representing the shortest distance between the watermarked product and the "not watermarked" region.

If the watermark is made up of many bits and embedded by a quantization method, the watermarked signal is located in a hyper-cube; similar attempts crossing the faces of the hyper-cube can provide adequate information about the quantizers. This way, the attacker can also remove or change the embedded watermark without causing serious distortion.

B. Secure Watermarking Algorithm

A secure watermarking scheme containing a tailor-made embedding algorithm, and a corresponding detection method is designed against the sensitivity attack. The result of watermark detection is also binary, "watermarked" or "not watermarked."

The embedding procedure is as follows.

1) Decompose a host image into three levels with 2-D orthogonal wavelet transform, denoting the coefficients $C_{l,\theta}(p,q)$, where l and θ indicate the levels and orientations, respectively (see Fig. 1).
2) Generate $(M+1)$ data-groups in a pseudo-random manner, where M is a system parameter. The number of elements in each data-group is equal to that of the host DWT coefficients, and all elements in the groups, $S_{l,\theta}^{(m)}(p,q)(0 \leq m \leq M)$, are mutually independent and satisfy a standard Gaussian distribution with zero mean and unit standard deviation. Here, the $(M+1)$ data-groups, or the way of generation, are shared between the watermark inserter and a corresponding detector but kept confidential to any third parties.
3) The watermark is embedded into the DWT coefficients. In order to improve robustness, the subbands that are more perceptually sensitive to noise, i.e., the nondiagonal high-level subbands, should carry more watermark components, since important subbands are less affected by compression coding. In [9], an empirical model of perceptual tolerance of coefficients in different subbands to noise is formulated based on HVS experiments, which is expressed as a product of two factors depending on DWT level and orientation

$$W_{l,\theta} = \left\{ \begin{array}{ll} 1.00, & \text{if } l=0 \\ 0.32, & \text{if } l=1 \\ 0.16, & \text{if } l=2 \end{array} \right\} \times \left\{ \begin{array}{ll} \sqrt{2}, & \text{if } \theta = 1 \\ 1, & \text{otherwise} \end{array} \right\}. \quad (2)$$

For instance, $W_{0,1} = 1.414$, and $W_{2,3} = 0.16$. A smaller $W_{l,\theta}$ corresponds to higher perceptual sensitivity. Calculate

$$T_{l,\theta}^{(m)}(p,q) = \frac{S_{l,\theta}^{(m)}(p,q)}{W_{l,\theta}} \Bigg/ \sqrt{\sum_l \sum_\theta \left(\frac{N_{l,\theta}}{W_{l,\theta}^2} \right)} \quad m = 0,1,\ldots,M \quad (3)$$

where $N_{l,\theta}$ are the numbers of coefficients in different subbands. Thus, $T_{l,\theta}^{(m)}(p,q)$ in nondiagonal high-level subbands generally have larger absolute values. Modify the DWT coefficients

$$C'_{l,\theta}(p,q) = C_{l,\theta}(p,q) - \sum_{m=0}^{M} \left[\frac{u_m(C)}{A_m} \cdot T_{l,\theta}^{(m)}(p,q) \right] \quad (4)$$

where

$$A_m = \sum_l \sum_\theta \sum_p \sum_q \left[T_{l,\theta}^{(m)}(p,q) \right]^2, \quad m = 0,1,\ldots,M \quad (5)$$

and u_m are functions of $C_{l,\theta}(p,q)$

$$u_m(C) = \sum_l \sum_\theta \sum_p \sum_q \left\{ \left[C_{l,\theta}(p,q) - \overline{C}_{l,\theta} \right] \cdot T_{l,\theta}^{(m)}(p,q) \right\} \quad m = 0,1,\ldots,M. \quad (6)$$

In (6), $\overline{C}_{l,\theta}$ is an average of the coefficients in subband (l,θ). This way, the coefficients in nondiagonal high-level subbands are used to carry more watermark components.

4) All modified DWT coefficients are inversely transformed to yield a watermarked copy.

According to the central limit theorem, A_m and u_m are approximately Gaussian. With a large size of host image, A_m is very close to its mean 1, while the mean of u_m is 0, and its standard deviation

$$\sigma_u(C) = \sqrt{\frac{\sum_l \sum_\theta \sum_p \sum_q \left\{ \left[C_{l,\theta}(p,q) - \overline{C}_{l,\theta} \right] / W_{l,\theta} \right\}^2}{\sum_l \sum_\theta \left(N_{l,\theta} / W_{l,\theta}^2 \right)}}. \quad (7)$$

So, PSNR due to watermark embedding is

$$\text{PSNR}_\text{w} \approx 10 \cdot \log_{10} \left[\frac{255^2 \cdot N_1 \cdot N_2}{\sigma_u^2 \cdot (M+1)} \right] \quad (8)$$

where N_1 and N_2 are, respectively, the numbers of rows and columns in the host image. Although sensitive coefficients are more affected, a small system parameter M can be assigned to ensure imperceptibility. Since the elements in the data-groups are mutually independent, the $(M+1)$ data-groups $T^{(m)}$, treated as vectors, are approximately orthogonal to each other. In other words, the relationship between different u_m is very weak so that all u_m in a watermarked copy are approximately zero.

The following is the procedure for blind watermark-detection.

1) Denote the DWT coefficients of an input image $C''_{l,\theta}(p,q)$. After obtaining $C''_{l,\theta}(p,q)$, $W_{l,\theta}$, and $T_{l,\theta}^{(m)}(p,q)$ in the

Fig. 2. Watermarked image with PSNR = 39.5 dB.

same way, calculate the $(M+1)$ values of $u_m(C'')$ using C'' and $T^{(m)}$ as in (6) and

$$U = \sqrt{M} \cdot |u_0(C'')| - \sum_{m=1}^{M} |u_m(C'')|. \tag{9}$$

2) Calculate

$$E = 0.8 \cdot (M - \sqrt{M}) \cdot \sigma_u(C'') \tag{10}$$

where $\sigma_u(C'')$ is derived from $C''_{l,\theta}(p,q)$ and $W_{l,\theta}$ according to (7), and the detection function

$$F = U + E/2. \tag{11}$$

3) If $F > 0$, the detector outputs the result "watermarked," otherwise, "not watermarked."

If a digital product watermarked with the above-mentioned embedding scheme is sent to the detector, in other words, $C'' = C'$, all u_m will be close to zero; therefore, $F \approx E/2 > 0$. If the input product contains no watermark, since each $u_m(C'')$ is Gaussian with zero mean and a standard deviation $\sigma_u(C'')$, according to the central limit theorem, U is approximately Gaussian with a mean

$$(\sqrt{M} - M) \cdot \sigma_u(C'') \int_{-\infty}^{+\infty} \frac{|t|}{\sqrt{2\pi}} e^{-t^2/2} \mathrm{d}t \approx 0.8 \cdot (\sqrt{M} - M) \cdot \sigma_u(C'') = -E \tag{12}$$

and a standard deviation as in (13), shown at the bottom of the page. Thus, the mean of F is $-E/2$, and a value in the middle of $[E/2, -E/2]$, i.e., 0, is used as the detection threshold. In this case, probability of incorrectly making a "watermarked" decision for an original signal is

$$\begin{aligned} \mathrm{P_e} &= \mathrm{P}(F > 0) = \mathrm{P}\left(U > -\frac{E}{2}\right) \\ &= \int_{-\frac{E}{2}}^{+\infty} \frac{1}{\sqrt{2\pi}\sigma_U} \exp\left[-\frac{(t+E)^2}{2 \cdot \sigma_U^2}\right] \mathrm{d}t. \end{aligned} \tag{14}$$

Letting

$$v = (t + E)/\sigma_U \tag{15}$$

(14) can be rewritten as

$$\mathrm{P_e} = \int_{\frac{E}{2\sigma_U}}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-v^2/2} \mathrm{d}v = \int_{0.47(\sqrt{M}-1)}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-v^2/2} \mathrm{d}v. \tag{16}$$

Obviously, a large M means small $\mathrm{P_e}$. In order to ensure an error probability less than 10^{-9}, M must not be less than 189. To avoid excessive distortion, $M = 189$ is recommended.

C. Security Analysis

Security of the proposed watermarking scheme relies on the confidentiality of the data-groups $T^{(m)}$. The watermark detector is a black box to any attacker who may know $W_{l,\theta}$, σ_u, and the detection output of any input signal but does not know the values of u_m. Therefore, it is impossible for him/her to estimate $T^{(m)}$ by measuring the sensitivities of u_m to the modifications in different directions. In other words, the attacker cannot modify the values of u_m by simply adding/subtracting $T^{(m)}$ to/from a watermarked product to flip the detection result.

On the other hand, with a product watermarked using the described method, an attacker can still obtain $\mathbf{v}_\mathrm{A}$ using sensitivity attack, that is, a vector representing the shortest distance from a constructed signal near the boundary to another region corresponding to an opposite detection result. However, this cannot be used to move a watermarked product away from the "watermarked" region with low distortion. In other words, adding/subtracting $\mathbf{v}_\mathrm{A}$ to/from a watermarked product does not destroy the embedded watermark. The key of the proposed technique is that sensitivity of the detection function to the data-group $T^{(0)}$ is $\sqrt{M}$ times greater than sensitivities to other M data-groups $T^{(1)}, T^{(2)}, \ldots, T^{(M)}$ so that the detection function of an attacked product is rebounded to $E/2$, leading to a robust "watermarked" decision.

In the sensitivity attack, when an attacker possesses a watermarked copy and an available detector, he can construct a signal near the decision boundary by introducing a strong noise. This means that the detection function of the disturbed signal is close to zero. Then, the attacker measures the sensitivity of detection

$$\sigma_U = \sqrt{2M\sigma_u(C'') \left[\int_{-\infty}^{+\infty} \frac{t^2}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} \mathrm{d}t - \left(\int_{-\infty}^{+\infty} \frac{|t|}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} \mathrm{d}t \right)^2 \right]} \approx 0.85\sqrt{M} \cdot \sigma_u(C'') \tag{13}$$

TABLE I
VALUES OF DETECTION FUNCTION AND PSNR IN THE COMPRESSED IMAGES WITH DIFFERENT QUALITY FACTORS

JPEG quality factor	70	40	20	10
PSNR (dB)	32.5	29.1	26.9	25.1
Detection function	9.6×10^3	6.1×10^3	2.0×10^3	-2.8×10^3

function to N orthogonal vectors by using the detector to produce a signal $\mathbf{v}_{\mathrm{A}}$ according to (1). As mentioned above, $\mathbf{v}_{\mathrm{A}}$ is roughly proportional to the gradient of the detection function F. Denote DWT coefficients of the disturbed signal $D_{l,\theta}(p,q)$ and that of the gradient $G_{l,\theta}(p,q)$. According to (9), adding $T^{(0)}$ to D will result in an increase of $u_0(D)$ by 1 and an increase of U by $\mathrm{sign}[u_0(D)]\cdot\sqrt{M}$, while adding $T^{(m)}(m=1,2,\ldots,M)$ to D will result in an increase of $u_m(D)$ by 1 and an increase of U by $-\mathrm{sign}[u_m(D)]$, where the operation $\mathrm{sign}(\cdot)$ returns $+1$ or -1 according to the sign of the argument. The case that $u_m(D)(m=0,1,2,\ldots,M)$ equals zero is ignored since its probability is very small. On the other hand, modification of E derived from adding/subtracting the data-groups $T^{(m)}$ or increasing/decreasing the value of $D_{l,\theta}(p,q)$ and pixels is significantly smaller. Thus, the DWT coefficients of the detection function's gradient

$$G_{l,\theta}(p,q)\approx\sqrt{M}\cdot\mathrm{sign}\left[u_0(D)\right]\cdot T^{(0)}_{l,\theta}(p,q)-\sum_{m=1}^{M}\left\{\mathrm{sign}\left[u_m(D)\right]\cdot T^{(m)}_{l,\theta}(p,q)\right\}. \quad (17)$$

Although gradient of the detection function is dependent on the disturbed signal, the absolute values of $u_m(G)$ are constant

$$|u_0(G)|\approx\sqrt{M};\quad\text{and}\quad|u_m(G)|\approx1,\quad m=1,2,\ldots,M. \quad (18)$$

Since $\mathbf{v}_{\mathrm{A}}$ is approximately proportional to the gradient, the DWT coefficients of $\mathbf{v}_{\mathrm{A}}$ are also approximately proportional to the DWT coefficients of the gradient, $G_{l,\theta}(p,q)$.

Then, the attacker adds $\mathbf{v}_{\mathrm{A}}$ to the watermarked signal with a certain strength to create an attacked signal. Denoting DWT coefficients of the attacked signal as $A_{l,\theta}(p,q)$

$$A_{l,\theta}(p,q)=C'_{l,\theta}(p,q)+\alpha\cdot G_{l,\theta}(p,q) \quad (19)$$

thus

$$\begin{aligned}u_0(A)&=u_0(C')+\alpha\cdot u_0(G)\\&=\alpha\cdot\sqrt{M}\cdot\mathrm{sign}\left[u_0(D)\right] \quad (20)\\u_m(A)&=u_m(C')+\alpha\cdot u_m(G)\\&=-\alpha\cdot\mathrm{sign}\left[u_m(D)\right],\quad m=1,2,\ldots,M. \quad (21)\end{aligned}$$

According to (9), $U=0$. In fact, sensitivity of U to the data-group $T^{(0)}$ is $\sqrt{M}$ times greater than sensitivities to other M data-groups so that the aggregate effects of modifications to U counteracts. Since modification of E due to adding/subtracting the data-groups $T^{(m)}$ is small, the detection function of the attacked signal is still approximately $E/2$, always greater than 0. This means the sensitivity attack cannot remove the embedded watermark.

III. EXPERIMENTAL RESULTS

A 960×1280 still image captured by a digital camera was used as the original test signal. The system parameter $m=189$ was chosen and Haar transform used. Fig. 2 shows the watermarked image with PSNR 39.5 dB. Values of the detection function F for the original and watermarked images were -1.7×10^4 and 1.5×10^4, respectively.

When compressing the watermarked image into JPEG with quality factors above 20 (PSNR $>$ 26.9 dB), the detection function was always positive, indicating that the compressed images contain a watermark. Table I presents the F values and PSNR of the compressed images. When adding white Gaussian noise to the watermarked image, the watermark could not be erased as long as PSNR is greater than 18 dB.

On the other hand, after obtaining a signal $\mathbf{v}_{\mathrm{A}}$ using sensitivity attack and subtracting it from the watermarked image with PSNR 24 dB, the values of detection function was 2.5×10^3. The embedded watermark could not be destroyed unless the attack was so strong that PSNR became less than 14 dB, which would make the image worthless.

IV. CONCLUSION

The described watermarking scheme contains a tailor-made embedding algorithm and a corresponding detection method and is capable of resisting sensitivity attacks. Adding/subtracting a "watermark" signal estimated in a sensitivity attack to/from a watermarked product does not affect the detection function defined in the proposed technique so that the embedded watermark is not removable and the watermark detector can safely be distributed as a black box for unlimited use.

REFERENCES

[1] F. A. P. Petitcolas, R. J. Anderson, and M. G. Kuhn, "Information hiding—a survey," *Proc. IEEE*, vol. 87, no. 7, pp. 1062–1078, Jul. 1999.

[2] S. Voloshynovskiy, S. Pereira, T. Pun, J. Eggers, and J. Su, "Attacks on digital watermarking: classification, estimation-based attacks, and benchmarks," *IEEE Commun. Mag.*, vol. 39, no. 8, pp. 118–126, Aug. 2001.

[3] I. J. Cox and J. P. M. G. Linnartz, "Some general methods for tampering with watermarks," *IEEE J. Sel. Areas Commun.*, vol. 16, no. 4, pp. 587–593, May 1998.

[4] J.-P. Linnartz and M. van Dijk, "Analysis of the sensitivity attack against electronic watermarks in images," Proc. 2nd Int. Workshop Information Hiding, Lecture Notes in Computer Science, vol. 1525, Springer, 1998, pp. 258–272.

[5] M. F. Mansour and A. H. Tewfik, "LMS-based attack on watermark public detectors," in *Proc. IEEE Int. Conf. Image Processing*, 2002, vol. 3, pp. 649–652.

[6] P. Comesaña, L. Pérez-Freire, and F. Pérez-González, "The return of the sensitivity attack," Proc. 4th Int. Workshop Digital Watermarking, Lecture Notes in Computer Science, vol. 3710, Springer, 2005, pp. 260–274.

[7] T. Kalker, J.-P. Linnartz, and M. van Dijk, "Watermark estimation through detector analysis," in *Proc. IEEE Int. Conf. Image Processing*, Chicago, IL, Oct. 1998, vol. 1, pp. 425–429.

[8] I. Venturini, "Oracle attacks and covert channels," Proc. 4th Int. Workshop Digital Watermarking, Lecture Notes in Computer Science, vol. 3710, Springer, 2005, pp. 171–185.

[9] A. S. Lewis and G. Knowles, "Image compression using the 2-D wavelet transform," *IEEE Trans. Image Process.*, vol. 1, no. 2, pp. 244–250, Feb. 1992.