ELSEVIER

Signal Processing 82 (2002) 1801–1804

SIGNAL PROCESSING

www.elsevier.com/locate/sigpro

Watermarking scheme capable of resisting attacks based on availability of inserter

Xinpeng Zhang*, Shuozhong Wang

School of Communication and Information Engineering, Shanghai University, 149 Yanchang Road, Shanghai, 200072, People's Republic of China

Received 29 June 2002

Abstract

Attacks based on the presence of watermark inserter are easy to perform since they make use of similarity between an original watermark and additionally added ones by using the same inserter and key. In this paper, a novel watermarking scheme capable of resisting inserter attacks is proposed. Watermark signals corresponding to the same key are mutually independent if they are randomly selected using the described technique. Thus the inserter attack is invalidated. Performance of the proposed method is studied, and simulation experiments presented.

Keywords: Digital watermarking; Attack; Inserter

1. Introduction

For the protection of intellectual property rights, watermarks can be embedded imperceptibly into multimedia data [2,5]. In the meantime, various attacks attempting to invalidate the embedded watermarks have also emerged. One type of malicious attacks is to remove or nullify the embedded watermark by using an available detector or inserter [1,3]. This is possible because, for the convenience of users, the copyright owner may need to distribute his watermark detector and the key. Moreover, the inserter may also be available to the public if the system is widely in use.

The attack strategy using watermark inserter is based on the difference between the original watermarked signal and a derived copy generated by embedding a second watermark on top of the original one. The maliciously added secondary watermark is identical or sufficiently similar to the original.

Assume that a watermark $\mathbf{W}$ is embedded into the host signal $\mathbf{I}$ using the inserter E and a key k to produce a watermarked signal $\mathbf{I}'$,

$$\mathbf{I}' = E(\mathbf{I}, \mathbf{W}, k). \tag{1}$$

For the convenience of detection, the embedded $\mathbf{W}$ and the key should be available to consumers. If $\mathbf{W}$ is a pseudo-random signal determined by k, only the key is needed in detection. Practically the key is also available to attackers. When an attacker has access to the inserter, he can embed an additional watermark similar to the original one into the watermarked product

$$\mathbf{I}'' = E(\mathbf{I}', \mathbf{W}, k). \tag{2}$$

If $\mathbf{W}$ is embedded by a method of addition without using perceptual model, the original and additional

* Corresponding author. Tel.: +86-21-5633-1964; fax: +86-21-5633-6908.

E-mail addresses: zhangxinpeng@263.net (X. Zhang), shuowang@yc.shu.edu.cn (S. Wang).

0165-1684/02/$ - see front matter
PII: S0165-1684(02)00344-4

watermarks should be identical

$$\mathbf{I}'' - \mathbf{I}' = \mathbf{I}' - \mathbf{I}. \tag{3}$$

If a perceptual model related to the host data is used in embedding, these two marks will not be identical but similar enough to each other. In this case the equality sign in (3) should be replaced with a sign of approximation. With both $\mathbf{I}'$ and $\mathbf{I}''$ at hand, the attacker is able to produce an illegal product, $2\mathbf{I}' - \mathbf{I}''$, that is free of any watermark. In other words, the watermark is removed.

A novel watermarking scheme against inserter attacks is proposed in this paper, in which mutually independent watermark signals are used.

2. Watermarking scheme and performance analysis

The embedding procedure is as follows:

1. The host data used for embedding are first re-organized using techniques such as linear transforms, data grouping or re-mapping, etc. DC coefficients are removed and proper embedding positions selected. This leads to a vector having a zero mean, $\mathbf{C} = C(0), C(1), \ldots, C(N-1)$.
2. A key $k = (k_0, k_1)$ is chosen by the copyright owner, where k_0 is an arbitrary integer, and k_1 is an integer within the interval $[N/3, 2N/3]$ and is prime to N. Define
$$f(i) = (k_0 + k_1 i) \bmod N, \quad i = 0, 1, \ldots, N-1 \tag{4}$$
Clearly, a one-to-one mapping between i and $f(i)$ exists.
3. An i.i.d. watermark signal, $\mathbf{W} = W(0), W(1), \ldots, W(N-1)$, with a zero mean and standard deviation σ_w is randomly selected.
4. The watermark is embedded into the host data as follows.
$$C'(i) = C(i) + W(i) + W[f^{-1}(i)], \quad i = 0, 1, \ldots, N-1. \tag{5}$$
Here, the watermark data and their permuted version are added to the host to produce the marked coefficients. After embedding, the watermark signal, $\mathbf{W}$, is abandoned.

In the detector, only the key $k = (k_0, k_1)$ is used, and both the host data and $\mathbf{W}$ are not needed. The following correlation parameter is first calculated

$$R = \frac{1}{N}\sum_{i=1}^{N} C'[f(i)]C'(i). \tag{6}$$

Then, the mean value of R can be obtained

$$E(R) = E\left\{\frac{1}{N}\sum_{i=1}^{N}\{C[f(i)] + W[f(i)] + W(i)\} \times \{C(i) + W(i) + W[f^{-1}(i)]\}\right\}. \tag{7}$$

Because of the finite distance between i and $f(i)$, correlation between $C(i)$ and $C[f(i)]$ is negligible. Since $C(i)$ is zero mean,

$$E\left\{\frac{1}{N}\sum_{i=1}^{N} C[f(i)]C(i)\right\} = 0. \tag{8}$$

In other words, the mean of R is zero if the data are "not watermarked". Moreover, $\mathbf{W}$ is independent of $\mathbf{C}$, then

$$\begin{aligned} &E\left\{\frac{1}{N}\sum_{i=1}^{N} C(i)W(i)\right\} \\ &= E\left\{\frac{1}{N}\sum_{i=1}^{N} C[f(i)]W(i)\right\} \\ &= E\left\{\frac{1}{N}\sum_{i=1}^{N} C(i)W[f(i)]\right\} = 0. \end{aligned} \tag{9}$$

And because $W(i)$s are mutually independent,

$$\begin{aligned} &E\left\{\frac{1}{N}\sum_{i=1}^{N} W[f(i)]W(i)\right\} \\ &= E\left\{\frac{1}{N}\sum_{i=1}^{N} W(i)W[f^{-1}(i)]\right\} = 0. \end{aligned} \tag{10}$$

Therefore, the expectation of R equals the energy contained in the watermark

$$E(R) = \sigma_w^2. \tag{11}$$

So, a threshold is chosen at the center between 0 and σ_w^2: if R is greater than $\sigma_w^2/2$, a "watermarked" decision is made, otherwise "not watermarked".

When an inserter attack is attempted, a secondary watermark $\mathbf{W}'$, independent of the original $\mathbf{W}$, is also randomly selected. Thus, the coefficient corresponding to $2\mathbf{I}' - \mathbf{I}''$ is

$$C''(i) = C(i) + W(i) + W[f^{-1}(i)] - W'(i) - W'[f^{-1}(i)], \quad i = 0, 1, \ldots, N-1. \tag{12}$$

Due to the independence between $\mathbf{W}$ and $\mathbf{W}'$, the introduction of $\mathbf{W}'$ does not decrease R, but, on the contrary, increases it so that the mean of the parameter R is doubled

$$E(R) = E\left\{\frac{1}{N}\sum_{i=1}^{N} C''[f(i)]C''(i)\right\} = 2\sigma_w^2. \tag{13}$$

The increase of $E(R)$ not only maintains the "watermarked" decision, but also provides a clue that the product has been tampered by an inserter attacker.

An attacker can always succeed since inserter attacks can be made repeatedly. Therefore, it is essential to make the attack too costly to be effective. Assume that σ_c is the standard deviation of the host data, $\mathbf{C}$. Standard deviation of $C(i)C[f(i)]$, denoted σ_{c^2}, equals variance of $C(i)$:

$$\sigma_{c^2} = \sqrt{\int\int C^2[f(i)]C^2(i)\mathrm{d}C[f(i)]\mathrm{d}C(i)} = \sqrt{\int C^2[f(i)]\mathrm{d}C[f(i)] \int C^2(i)\mathrm{d}C(i)} = \sigma_c^2. \tag{14}$$

If a watermark is absent, R is Gaussian according to the central limit theorem,

$$R \sim N\left(0, \frac{\sigma_c^2}{\sqrt{N}}\right). \tag{15}$$

If the data is watermarked and $\mathbf{W}$ is white Gaussian, then

$$R \sim N\left(\sigma_w^2, \frac{\sqrt{\sigma_c^4 + 4\sigma_c^2\sigma_w^2 + 5\sigma_w^4}}{\sqrt{N}}\right). \tag{16}$$

In general, standard deviation of the watermark signal, σ_w, is significantly smaller than that of the host data, σ_c. From (16), the distribution of R is approximately

$$R \sim N\left(\sigma_w^2, \frac{\sigma_c^2}{\sqrt{N}}\right). \tag{17}$$

With a threshold at $\sigma_w^2/2$, a decision very close to the maximum likelihood estimation is made and the probability of false decision can be calculated

$$P = 1 - \Phi\left(\frac{\sigma_w^2\sqrt{N}}{2\sigma_c^2}\right) = \frac{1}{\sqrt{2\pi}}\int_{\sigma_w^2\sqrt{N}/2\sigma_c^2}^{\infty} \exp\left(-\frac{1}{2}t^2\right)\mathrm{d}t. \tag{18}$$

After the inserter attack,

$$R \sim N\left(2\sigma_w^2, \frac{\sqrt{\sigma_c^4 + 8\sigma_c^2\sigma_w^2 + 18\sigma_w^4}}{\sqrt{N}}\right). \tag{19}$$

In this case, probability of false rejection becomes

$$P' = 1 - \Phi\left(\frac{3\sigma_w^2\sqrt{N}}{2\sigma_c^2}\right). \tag{20}$$

Obviously, the average number of attempts for a successful attack is $1/P'$. With a large N, the value of $1/P'$ is large enough so that an effective attack is practically impossible.

3. Simulation experiments

As a watermarking scheme against inserter attacks, the proposed technique can be applied to different digital media such as image, audio, and video. Also, it is not restricted to embedding locations and any specific transform used. In the present study, nonetheless, experiments were carried out on still images using a DCT technique to embed watermark into a lower-middle band in the transform domain.

The first experiment is to check for the conclusion drawn in the above performance analysis. A 512×512 test image Lena was segmented into $64 \times 64 = 4096$ blocks, each sized 8×8. Two-dimensional DCT was then performed on the blocks. A 4096×1 vector, to be used for hiding the watermark signal, is formed with the lower-middle frequency coefficients at position $(2, 2)$ in each 8×8 block. The computed standard

deviation of these coefficients was $\sigma_c = 24.07$. Selecting $k = (0, 1759)$ and $\sigma_w = 0.16 \times \sigma_c = 3.85$, a random watermark signal was embedded using the proposed method. The peak signal-to-watermark-power ratio (PSWR) is 54.5 dB. From (19), $R \sim N(29.66, 9.94)$, so $1/P' = 79$. The test image was watermarked to produce 50 differently watermarked images, each containing an independent watermark signal, and was attacked repeatedly using the inserter. To remove the watermark, the average number of attempts needed for each image was 75, which is close to the theoretically calculated 79.

In the second experiment, the watermarking scheme was modified to improve robustness. After block DCT on the test image, four coefficients at $(2,2)$, $(2,3)$, $(3,2)$ and $(3,3)$ in each block were chosen. All the $4096 \times 4 = 16\,384$ data were then re-arranged into a 128×128 matrix after a pseudo-random shuffling. The purpose of shuffling was to remove unwanted correlation between neighboring coefficients so that a uniform expected energy in the coefficients would be obtained after further transformation. The particular way of shuffling could also be used as part of the key. A second-layer DCT was then performed on the shuffled data to produce a 16384×1 vector, which was used as the host data to hide the watermark. The calculated standard deviation of these coefficients was $\sigma_c = 17.55$. Selecting $k = (0, 10123)$ and $\sigma_w = 0.225 \times \sigma_c = 3.95$, a random watermark signal was finally embedded into the host. The embedded watermark was highly invisible to subjective viewing. In the objective assessment, the calculated PSWR of the watermarked image was 48.2 dB, much higher than the lower bound of 38 dB for the objective criterion of invisibility in watermark embedding as suggested in [4].

Using (15) and (16), the probability of a "not watermarked" decision for a marked images was 1.6×10^{-3}, while the probability of a "watermarked" decision for an image free of watermark was 5.9×10^{-4}. Finally, it can be shown that the watermarking technique is capable of resisting inserter attacks as $P' = 1 - \Phi(8.069) = 3.3 \times 10^{-16}$ is extremely small, which means that, with a realistic number of attempts, a successful inserter attack is impossible.

4. Conclusion

In conclusion, the key step of the proposed approach is to randomly select watermark signals, which are mutually independent even derived from a single key. Because of the independence, the subtraction of another watermark derived from a same key cannot remove the original mark. Further, it will increase the correlation parameter R to give a warning that the product has been attacked with the inserter. Therefore the inserter attack is invalidated.

Vulnerability of the watermark mainly depends on such factors as the choice of $C(i)$, i.e., the type of transform, the length of coefficients, the chosen spectral position, etc., and the embedding strength (the magnitude of $W(i)$). Generally speaking, with a particular transform, a robust watermark requires that the coefficient sequence is not too short, and they should be chosen from relatively low spectral positions. Meanwhile, $W(i)$ should be as large as possible within the constraint of imperceptibility.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 60072030) and Key Disciplinary Development Program of Shanghai.

References

[1] I.J. Cox, J.P.M.G. Linnartz, Some general methods for tampering with watermarks, IEEE J. Select. Areas Commun. 16 (1998) 587–593.
[2] F. Hartung, M. Kutter, Multimedia watermarking techniques, Proc. IEEE 87 (1999) 1079–1107.
[3] T. Kalker, J. Linnartz, M. Dijk, Watermark estimation through detector analysis, Proceedings of the IEEE International Conference on Images Processing, Vol. 1, Chicago, IL, USA, October 1998, pp. 425–429.
[4] F.A.P. Petitcolas, R.J. Anderson, Evaluation of copyright marking systems, Proceedings of the IEEE Multimedia Systems, Florence, Italy, June 1999, pp. 574–579.
[5] F.A.P. Petitcolas, R.J. Anderson, M.G. Kuhn, Information hiding—a survey, Proc. IEEE 87 (1999) 1062–1078.