Near-Perfect Cover Image Recovery Anti-Multiple Watermark Embedding Approaches

Chao-Yong Hsu
Institute of Information Science
Academia Sinica
Taipei, Taiwan 115, ROC
Email: cyhsu@iis.sinica.edu.tw

Chun-Shien Lu
Institute of Information Science
Academia Sinica
Taipei, Taiwan 115, ROC
Email: lcs@iis.sinica.edu.tw

Abstract— **Robustness is a critical requirement for a watermarking scheme to be practical. Especially, in order to resist geometric distortions a common way is to locally insert multiple-redundant watermarks in the hope that partial watermarks could still be detected. However, there exist watermark-estimation attack (WEA), such as the collusion attack, that can remove watermarks while making the attacked data further transparent to its original. Another kind of attack is the copy attack, which can cause protocol ambiguity within a watermarking system. The aim of this paper is to propose an efficient cover data recovery attack, which is more powerful than the conventional collusion attack. To this end, we begin by gaining insight into the WEA, leading to formal definitions of optimal watermark estimation and near-perfect cover data recovery. Subject to these definitions, an exquisite collusion attack is derived. Experimental results verify the effectiveness of the proposed watermark estimation and recovery algorithm.**

I. Introduction

Robustness is known to be a critical issue affecting the practicability of a watermarking system. In the literature, robustness is usually examined with respect to removal attacks or geometrical attacks or both. However, there indeed exist attacks that can defeat a watermarking system without sacrificing media quality. In particular, the collusion attack [10], [11], which is a removal attack, and the copy attack [4], which is a protocol attack, are typical examples of attacks that can achieve the aforementioned goal. The common step used to realize a collusion or copy attack is watermark estimation, which is easily accomplished by means of denoising. Consequently, we call both the collusion attack and copy attack watermark-estimation attacks (WEAs) [5]. In this study, we particularly focus on the collusion attack.

The aim of the collusion attack is to collect and analyze a set of watermarked media data so that an unwatermarked copy can be constructed to create the false negative problem. In digital watermarking, a collusion attack naturally occurs in video watermarking because a video is composed of many frames, and one way of watermarking a video is to embed the same watermark into all the frames. This scenario was first addressed in [11]. However, we argue that [5] the collusion attack is not exclusively applied to video watermarking. In the past few years, image watermarking with resistance to geometrical attacks has received much attention because even a slight geometrical distortion may disorder the hidden watermark bits and disable watermark detection. In view of this fact, some researches [2], [12], [9], [14] inserted multiple redundant watermarks into an image in the hope that robustness could be maintained as long as partial watermarks existed. Commonly, various kinds of image units, such as blocks [14], meshes [2], or disks [12], are extracted as carriers for embedding. Taking advantage of this unique characteristic, we propose to treat each image unit in an image like a frame in a video; in this way, collusion attacks can be equally applied to those image watermarking methods that employ a multiple redundant watermark embedding strategy. Therefore, once the hidden watermarks are successfully removed by means of a collusion attack, the false negative problem occurs even though no geometrical attack is imposed on stego images. Of particular interest are possible fidelity improvements of attacked images as a result of a collusion attack.

When the hidden watermark is estimated and removed by means of collusion, it is necessary to check the presence or absence of a watermark. A simple way is to calculate a correlation (e.g., cross-correlation) and compare it against a threshold to make the final decision about the existence of a watermark. However, one may argue that this does not imply that the hidden watermark has been "optimally estimated and removed" by means of such a simple cross-correlation. This is because an "optimal" watermark detector [1], [8], which is usually based on exploiting the statistic characteristic of a host content, may be able to discover the hidden watermark. In order to address this issue, we don't evaluate the optimal estimation/removal of a watermark from the viewpoint of a watermark detector. On the contrary, we investigate how an embedded watermark could be "sufficiently" estimated/removed. In this paper, we propose a new watermark estimation and cover data recovery method. The comparison of our method with perceptual remodulation [13] is also evaluated.

II. How Watermark Could be Completely Removed?

Let $\mathbf{W}$ represent the original watermark with its energy extended by means of either a constant factor or a human visual system to enhance robustness. From an attacker's

0-7803-8834-8/05/$20.00 ©2005 IEEE.

perspective, the energy of each watermark value must be accurately predicted so that the previously added watermark energy can be completely subtracted to create an ideally unwatermarked image. If this goal could be achieved, it is said that watermark removal is effective without leaving sufficient residual watermark. Consequently, an estimated watermark's energy is closely related to the accuracy of the watermark removal attack. To better explain our point, several motivating scenarios are shown in Fig. 1, which illustrates the energy variations of (a) an original watermark; (b)/(d) an estimated watermark (illustrated in gray-scale); and (c)/(e) a residual watermark generated by subtracting the estimated watermark from the original watermark. We can observe from Fig. 1(a)$\sim$(c) that even though the watermark's sign bits are fully obtained (Fig. 1(b)), the residual watermark signal (Fig. 1(c)) still suffices to reveal the encoded message due to the original watermark's energies cannot be completely discarded. Furthermore, if the sign of an estimated watermark value is different from its original one, then any additional energy subtraction will not be helpful in improving removal efficiency. On the contrary, watermark removal in terms of energy subtraction operated in the opposite (wrong) polarity will undesirably damage the media data's fidelity. Actually, this corresponds to adding a watermark with higher energy into cover data without satisfying the masking constraint, as shown in Fig. 1(d). After Fig. 1(d) is subtracted from Fig. 1(a), the resultant residual watermark is illustrated in Fig. 1(e). By comparing Figs. 1(a) and (e), it is highly possible to reveal the existence of a watermark.

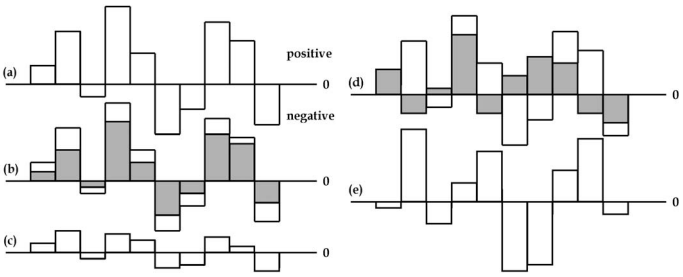

Fig. 1. Watermark estimation/removal illustrated with energy variation: (a) original embedded watermark with each white bar indicating the energy (determined using perceptual masking) of each watermark value; (b) gray bars show the energies of an estimated watermark with all the signs being the same as the originals (a); (c) the residual watermark obtained after removing the estimated watermark (b); (d) the energies of an estimated watermark with most the signs being opposite to those in (a); (e) the residual watermark derived from (d). In the above examples, sufficiently large correlations between (a) and (c), and between (a) and (e) exist, indicating the presence of a watermark.

The observations from Fig. 1 motivate us to formulate the definitions of "optimal watermark estimation" and "near-perfect cover data recovery." This implies that we try to recover a cover data from its stego version. If this goal can be achieved, even optimal watermark detector will fail to detect the hidden watermark; otherwise false positive will appear.

Definition 1 (Optimal Watermark Estimation): Given an original embedded watermark signal $\mathbf{W}$ and its approximate version $\mathbf{W^e}$ estimated from the stego image $\mathbf{X^s}$, the necessary condition for the optimal estimation of $\mathbf{W}$ as $\mathbf{W}^e$ is defined as

$$BER(sgn(\mathbf{W}), sgn(\mathbf{W^e})) \leq \tau, \tag{1}$$

where τ is watermarking algorithm- and application-dependent, and the sign function, $sgn(\cdot)$, is defined as

$$sgn(t) = \begin{cases} +1, & \text{if } t \geq 0, \\ -1, & \text{if } t < 0. \end{cases}$$

Basically, Definition 1 is naturally derived from Fig. 1 in that the "polarity of each watermark value is particularly crucial. To assist our later analysis, we use Θ to denote the set of indices satisfying $sgn(W^e(i)) = sgn(W(i))$ in Eq. (1). This is the first step, where the existence of a watermark may be efficiently eliminated if most sign bits of the watermark can be obtained by an attacker. Beyond this step, however, to avoid leaving a residual watermark (as illustrated in Fig. 1(c)) that can reveal the hidden watermark, accurate estimation of the energy of $\mathbf{W^e}$ is absolutely indispensable. In addition to Eq. (1), watermark removal can be completely achieved if the watermark energy to be subtracted is also larger than or equal to the added energy, i.e., $mag(W^e(i)) \geq mag(W(i))$, where $mag(t)$ denotes the magnitude $|t|$ of t. Therefore, it is said that $\mathbf{W^e}$ is an optimal estimation of $\mathbf{W}$ if and only if

$$\begin{aligned} &BER(sgn(\mathbf{W}), sgn(\mathbf{W^e})) \leq \tau \text{ and} \\ &mag(W^e(i)) \geq mag(W(i)) \ \ \forall i \in \Theta. \end{aligned} \tag{2}$$

Definition 2 (Near-Perfect Cover Data Recovery): Under the prerequisite that Definition 1 (Eq. (2)) is satisfied, it can be said that $\mathbf{X^r}$ is a near-perfect recovery of the cover image $\mathbf{X}$ if

$$PSNR(\mathbf{X}, \mathbf{X^r}) \approx \infty, \tag{3}$$

where $\mathbf{X^r} = \mathbf{X^s} - sgn(\mathbf{W^e})mag(\mathbf{W^e})$, $\mathbf{X^s} = \mathbf{X} + sgn(\mathbf{W})mag(\mathbf{W})$, and $sgn(\mathbf{v})$ and $mag(\mathbf{v})$ are two vectors representing the sign and magnitude of the elements in a vector $\mathbf{v}$, respectively.

It is noted that ideally Eq. (3) is satisfied only if $mag(\mathbf{W^e}) \approx mag(\mathbf{W})$; otherwise, even if the watermark values have been completely removed based on $mag(\mathbf{W^e}) >> mag(\mathbf{W})$, the quality of the attacked/recovered image would be undesirably degraded. Typically, evaluation of $mag(\mathbf{W^e})$ can be achieved by means of either averaging or remodulation. It should be noted that if the residual watermark (Fig. 1(c)) becomes empty or negatively correlated with the hidden watermark (Fig. 1(a)), then even optimal watermark detector is unable to detect the hidden watermark. Definition 2 has specified how a cover data could be recovered in a near-perfect manner. In Sec. III, a near-perfect cover data recovery algorithm will be described.

III. A Near-Perfect Cover Data Recovery Algorithm

"Near-perfect" here means that the hidden watermarks can be mostly removed with a high probability (say 90%) so that the recovered data is more similar to the cover data than the stego data. Under this circumstance, it is not necessary to worry about the detection ability of optimal watermark

detector; otherwise, they will run the risk of raising the false positive problem. Here, we shall take the block-based multiple self-reference watermarking method [14] as an example to explain the performance of our algorithm in removing the hidden watermarks. However, it should be noted that our algorithm can be extended to other methods [2], [9], [12] that adopt the similar concept of multiple redundant watermark embedding.

In the following, the method [14] where the watermark embedded in each image block is a bipolar sequence is briefly described. This watermark $\mathbf{W}$ is flipped and copied in each direction to produce a symmetric signal, which is repeated over the entire image. In the embedding process, both the expanded watermark signal and the cover image are first decomposed using wavelet transform. Then, the watermark signals are embedded into the cover image in the wavelet domain through linear additive modulation together with a perceptual masking model called "noise visibility function (NVF)" [14]. The NVF is basically a wavelet-based content-adaptive visual model so that the degree for each wavelet coefficient that can be modified without raising perceptual difference can be defined. Let $NVF_{k,l}(m,n)$ denote the masking threshold for the wavelet coefficient at the position (m,n) of subband k,l (where k denotes scale and l denotes orientation), and let $x_{k,l}(m,n)$ and $y_{k,l}(m,n)$ denote the cover and stego image wavelet coefficients, respectively. They are related as

$$\begin{aligned} y_{k,l}(m,n) &= x_{k,l}(m,n) \\ &+ (1-NVF_{k,l}(m,n)) \cdot w_{k,l}(m,n) \cdot S^e_{k,l} \\ &+ NVF_{k,l}(m,n) \cdot w_{k,l}(m,n) \cdot S^f_{k,l}, \end{aligned} \tag{4}$$

where $w_{k,l}(m,n)$'s denote the watermark wavelet coefficient, $S^e_{k,l}$ and $S^f_{k,l}$ denote the embedding strength for non-flat and flat regions, respectively.

Now, the proposed near-perfect cover data recovery algorithm based on the collusion estimation of watermark's signs and NVF-based estimation of watermark's magnitudes is described as follows. Let $\mathbf{W^e}$ be the watermark estimated by means of collusion. As pointed out in Fig. 1 and Definition 2, accurate estimation of watermark's magnitudes is crucial to completely remove the hidden watermarks. In fact, we would rather remove more watermark energy than it should be so that the watermark energy can be more guaranteed to be eliminated. Let $NVF^s_{k,l}(m,n)$ denote the masking threshold for a stego image. The wavelet coefficient for the recovered image $\mathbf{X^r}$ based on Definition 2 can be derived as

$$\begin{aligned} & z_{k,l}(m,n) = y_{k,l}(m,n) \\ - \; & [(1-NVF^s_{k,l}(m,n)) \cdot w^e_{k,l}(m,n) \cdot S^e_{k,l} \\ + \; & NVF^s_{k,l}(m,n) \cdot w^e_{k,l}(m,n) \cdot S^f_{k,l}](1+\epsilon_{k,l}(m,n)), \end{aligned} \tag{5}$$

where $w^e_{k,l}(m,n)$'s denote the estimated watermark wavelet coefficient and $\epsilon_{k,l}(m,n)$'s are used to more guarantee that the hidden watermark can be completely removed. By substituting Eq. (4) into Eq. (6) and assuming that the recovered image is equal to the cover image; i.e., $z_{k,l}(m,n) = x_{k,l}(m,n)$ for all k, l, i, and j, the desired parameters, $\epsilon_{k,l}(m,n)$'s, can be derived. To simplify analysis, we further assume that $\mathbf{W} = \mathbf{W^e}$; i.e., their watermark wavelet coefficients satisfy $w_{k,l}(m,n) = w^e_{k,l}(m,n)$ for all k, l, i,, and j. In this case, $\epsilon_{k,j}(m,n)$ can be ideally derived as

$$\epsilon_{k,l}(m,n) = \frac{NVF_{k,l}(m,n) - NVF^s_{k,l}(m,n)}{NVF^s_{k,l}(m,n) - \frac{S^e_{k,l}}{S^e_{k,l}-S^f_{k,l}}}. \tag{6}$$

In Eq. (6), $NVF_{k,l}(m,n)$'s are unknown since no cover data is available in a blind detection scenario to obtain its NVF. However, they can be approximately estimated if $\mathbf{X^s} - sgn(\mathbf{W^e})mag(\mathbf{W^e})$, as described in Definition 2, is used to obtain an approximate cover image. It should be noted that when $\epsilon_{k,l}(m,n)$'s are equal to zero, this algorithm degenerates to watermark remodulation [13]. If watermark's energy is estimated by means of averaging, this algorithm degenerates to conventional collusion attack.

In this section, we have derived how the watermark magnitude can be estimated to achieve complete watermark removal. In order to evaluate the performance of "near-perfect cover data recovery," it is best to compare the recovered image with the cover image to check how many watermark bits still survive in the recovered image.

IV. Experimental Results

In our experiments, ten varieties of standard cover images of size 512×512 were used for watermarking. In this study, Voloshynovskiy *et al.*'s block-based image watermarking approach [14] was chosen as the benchmark due to its strong robustness and computational simplicity. However, we would like to particularly emphasize that the proposed scheme is readily applied to other watermarking algorithms that implement the similar principle of embedding multiple redundant watermarks [2], [9], [12]. Wiener filter was used to perform denoising-based blind watermark extraction.

In order to verify how the hidden (content-independent) watermark could be removed by means of the proposed optimal watermark estimation algorithm (Sec. III), the survived watermark of the obtained recovered image was extracted using the cover image so that we can accurately check how many correct watermark bits still remain. Table I shows the BER values, which were obtained from comparing the original watermark and the extracted watermarks, and the PSNR values, which were calculated between the cover image and the recovered/stego image. As we can see from Table I that if $\epsilon_{k,l}(m,n) = 0$ is used, this corresponds to perceptual remodulation [13]. It is observed that PSNRs have been increased and most BERs fall into the interval between $50\% \sim 60\%$, which means that a significant part of watermark values is not completely removed.

However, if $\epsilon_{k,l}(m,n) \neq 0$ is adopted, BERs can be increased averagely as high as 0.9 except for some very smoothing images, which implies that our estimation and recovery algorithms are able to remove almost all the watermark

TABLE I

Validation of our estimation and recovery scheme. BER is computed between the original and the extracted watermarks. PSNR is computed between the cover and the recovered/stego images.

Stego image	$\mathbf{X_1^s}$	$\mathbf{X_2^s}$	$\mathbf{X_3^s}$	$\mathbf{X_4^s}$	$\mathbf{X_5^s}$	$\mathbf{X_6^s}$	$\mathbf{X_7^s}$	$\mathbf{X_8^s}$	$\mathbf{X_9^s}$	$\mathbf{X_{10}^s}$
PSNR (dB)	38.15	37.92	37.96	37.51	37.61	37.94	38.13	38.98	37.74	37.94
Recovered image using $\epsilon_{k,l}(m,n)=0$ [13]	$\mathbf{X_1^r}$	$\mathbf{X_2^r}$	$\mathbf{X_3^r}$	$\mathbf{X_4^r}$	$\mathbf{X_5^r}$	$\mathbf{X_6^r}$	$\mathbf{X_7^r}$	$\mathbf{X_8^r}$	$\mathbf{X_9^r}$	$\mathbf{X_{10}^r}$
PSNR (dB)	45.46	53.25	53.16	43.10	56.20	48.18	52.79	45.05	54.83	53.25
BER (%)	68.8	48.4	48.4	43.8	57.8	78.1	78.1	89.1	60.9	57.8
Recovered image using $\epsilon_{k,l}(m,n)\neq 0$	$\mathbf{X_1^r}$	$\mathbf{X_2^r}$	$\mathbf{X_3^r}$	$\mathbf{X_4^r}$	$\mathbf{X_5^r}$	$\mathbf{X_6^r}$	$\mathbf{X_7^r}$	$\mathbf{X_8^r}$	$\mathbf{X_9^r}$	$\mathbf{X_{10}^r}$
PSNR (dB)	53.15	55.30	54.38	59.65	58.28	53.44	53.31	49.06	56.08	54.23
BER (%)	82.1	85.7	85.1	89.6	89.4	90.5	86.3	80.5	87.8	84.0
Recovered image using $2\times\epsilon_{k,l}(m,n)$	$\mathbf{X_1^r}$	$\mathbf{X_2^r}$	$\mathbf{X_3^r}$	$\mathbf{X_4^r}$	$\mathbf{X_5^r}$	$\mathbf{X_6^r}$	$\mathbf{X_7^r}$	$\mathbf{X_8^r}$	$\mathbf{X_9^r}$	$\mathbf{X_{10}^r}$
PSNR (dB)	51.69	53.53	52.90	57.12	56.06	51.52	52.16	47.65	54.45	52.34
BER (%)	88.5	92.8	93.0	93.8	94.1	93.7	92.8	81.8	92.8	90.9

bits in a stego image. In addition, the obtained PSNRs are further improved than those obtained using [13] such that the recovered image can be more similar to its cover version. Since $\epsilon_{k,l}(m,n)$'s are approximated derived in Eq. (6), if $2\times\epsilon_{k,l}(m,n)$ is heuristically adopted, we show that (by comparing those results obtained using $\epsilon_{k,l}(m,n)=0$ and $\epsilon_{k,l}(m,n)\neq 0$) the BERs can be further increased and the PSNRs are moderate.

Under the circumstance that the proposed near-perfect cover data recovery algorithm is used, we are confident based on Table I that even the so-called "optimal watermark detector" [1], [8] is difficult to detect the survived (but few) watermark bits to sufficiently claim the existence of a watermark; otherwise, false positive probability is easy to occur. This validates our claim that efficient elimination of previously added watermark energy is indispensable to really remove the hidden watermark. In addition to the effective watermark removal, the recovered images were found to be more transparent to their cover ones than the stego images in terms of PSNR values. These experimental results demonstrate the performance of the proposed cover data recovery algorithm in defeating the multiple redundant watermark embedding methods that were originally addressed to tolerate geometric distortions.

V. Conclusion

Although multiple watermarks can be embedded into an image to withstand geometrical distortions, they are unfortunately vulnerable to collusion and copy attacks, and the desired geometric invariance is lost. In this study, we have proposed an efficient watermark estimation and recovery algorithm (which is regarded as an exquisite collusion attack) that can eliminate almost all watermark values. To cope with the watermark estimation attack (WEA), an anti-disclosure content-dependent watermark (CDW) with resistance to WEA has been investigated in [5]. In our recent paper [7], the proposed CDW has been combined with geometric-invariant image hash [3], [6] to obtain a mesh-based content-dependent watermarking scheme that can resist both the geometric and estimation attacks.

References

[1] M. Barni, F. Bartolini, A. E. Rosa, and A. Piva, "A New Decoder for the Optimum Recovery of Nonadditive Watermarks," *IEEE Trans. on Image Processing*, Vol. 10, No. 5, pp. 755-766, 2001.

[2] P. Bas, J. M. Chassery, and B. Macq, "Geometrically invariant watermarking using feature points," *IEEE Trans. Image Processing*, Vol. 11, No. 9, pp.1014-1028, 2002.

[3] C. Y. Hsu and C. S. Lu, "A Geometric-Resilient Image Hashing System and Its Application Scalability," *Proc. ACM Multimedia and Security Workshop*, pp. 81-92, Germany, 2004.

[4] M. Kutter, S. Voloshynovskiy, and A. Herrigel, "The Watermark Copy Attack," *Proc. SPIE: Security and Watermarking of Multimedia Contents II*, Vol. 3971, 2000.

[5] C. S. Lu and C. Y. Hsu, "Content-Dependent Anti-Disclosure Image Watermark," *Proc. 2nd Int. Workshop on Digital Watermarking (IWDW)*, LNCS 2939, pp. 61-776, Seoul, Korea, 2003.

[6] C. S. Lu, C. Y. Hsu, S. W. Sun, and P. C. Chang, "Robust Mesh-based Hashing for Copy Detection and Tracing of Images," *Proc. IEEE Int. Conf. on Multimedia and Expo: special session on Media Identification*, Taipei, Taiwan, 2004.

[7] C. S. Lu, S. W. Sun, and P. C. Chang, "Robust Mesh-based Content-dependent Image Watermarking with Resistance to Both Geometric Attack and Watermark-Estimation Attack," to appear in *Proc. SPIE: Security, Steganography, and Watermarking of Multimedia Contents VII (EI120)*, 2005.

[8] A. Nikolaidis and I. Pitas, "Asymptotically Optimal Detection for Additive Watermarking in the DCT and DWT Domains," *IEEE Trans. on Image Processing*, Vol. 12, No. 5, pp. 563-571, 2003.

[9] J. S. Seo and C. D. Yoo, "Localized image watermarking based on feature points of scale-space representation," *Pattern Recognition*, Vol. 37, pp. 1365-1375, 2004.

[10] K. Su, D. Kundur, D. Hatzinakos, "Statistical Invisibility for Collusion-resistant Digital Video Watermarking," *IEEE Trans. on Multimedia*, 2004.

[11] M. D. Swanson, B. Zhu, and A. H. Tewfik, "Multiresolution Scene-Based Video Watermarking Using Perceptual Models," *IEEE Journal on Selected Areas in Communications*, Vol. 16, No. 4, pp. 540-550, 1998.

[12] C. W. Tang and H. M. Hang, "A Feature-Based Robust Digital Image Watermarking Scheme," *IEEE Trans. Signal Processing*, Vol. 51, No. 4, pp.950-958, April 2003.

[13] S. Voloshynovskiy, S. Pereira, A. Herrigel, N. Baumgartner, and T. Pun, "Generalized Watermarking Attack Based on Watermark Estimation and Perceptual Remodulation," *Proc. SPIE: Security and Watermarking of Multimedia Contents II*, Vol. 3971, San Jose, CA, USA, 2000.

[14] S. Voloshynovskiy, F. Deguillaume, and T. Pun, "Multibit Digital Watermarking Robust against Local Nonlinear Geometrical Distortions," *Proc. IEEE Int. Conf. on Image Processing*, pp. 999-1002, 2001.