BSS: A New Approach for Watermark Attack

Jiang Du[*] Choong-Hoon Lee[*] Heung-Kyu Lee[*] Youngho Suh[**]
[*] Department of Computer Science
Korea Advanced Institute of Science and Technology
373-1 Guseong-Dong, Yuseong-Gu, Daejon, South Korea
[**] Content Technology Department, Computer & Software Technology Laboratory
Electronics and Telecommunications Research Institute
161 Gajeong-Dong, Yuseong-Gu, Daejeon, 305-350, South Korea
Email: [jdu, chlee, hklee] @casaturn.kaist.ac.kr syh@etri.re.kr

Abstract

Digital watermarking is the enabling technology to prove ownership on copyrighted material, detect originators of illegally made copies, monitor the usage of the copyrighted multimedia data, and analyze the spread spectrum of the data over networks and servers. Most watermarking methods for images and video can be viewed as a communications problem in which the watermark must be transmitted and received through a watermark channel. This channel includes distortions resulting from attacks and interference from the original digital data [1].

It is well accepted that an effective watermarking scheme may be described as the secure, imperceptible, robust communication of information by direct embedding in and retrieval from digital data. For verifying the security and robustness of watermarking algorithms, specific attacks have to be applied to test them. In this paper, using a theoretical approach based on random processes, signal processes, and communication theory, we propose a stochastic formulation of a new watermarking attack using blind source separation-based concept. The proposed attack consider the watermarking channel as a "black-box". A host image was passed through the "black-box", which include the watermarking embedding process, and then the watermarked image was produced. The watermarked image is viewed as linear mixtures of unknown source signals, and then we attempt to recover sources from their linear mixtures without resorting to any prior knowledge by using blind source separation theory.

1. Introduction

Digital watermarking systems can be viewed as digital communication systems. The information *b* is embedded directly and imperceptibly into digital multimedia documents, which is called "host documents" or "original documents", to form "marked signals". The embedded watermark should be reliably decodable even after further processing of the marked data, which is also denoted as attack against the embedded watermark. Such processing can be simple D/A-A/D conversion of the documents, but also a malicious attempt to impair watermark reception [2]. Here, digital watermarking is considered as the robust, imperceptible, and secure communication of information by embedding it in and retrieving it from other digital data. For instance, the information *b* can provide messages about the copyright holder of a document or indicate the copy state of the digital content. Over last years, many different watermarking schemes for large variety of data types have been developed. Potential applications of digital watermarking include copyright protection, distribution tracing, authentication, and conditional access control.

As the research area of digital watermarking matures, one can see some general trends in its development. There was initial work in the use of basic digital signal processing strategies for watermark embedding. Robustness-enhancing strategies were employed using intuition on human perception and basic communications. However, as the area has grown, some theory is emerging. This framework aims to unify much of the past work and establish technical insights for future algorithms. The new mathematical language for describing watermarking

Proceedings of the IEEE Fourth International Symposium on Multimedia Software Engineering (MSE'02)
2002114351/02 $17.00 © 2002 IEEE

IEEE COMPUTER SOCIETY

borrows tools from statistical communications and information theory [3]. Recently, more theoretical approaches have attempted to provide watermarking, and the larger field of information hiding, with a stronger foundation [4][5][6].

Many applications of digital watermarking, like ownership protection, copy/access control, and authentication, stay in rivalry environment where an adversary has incentives to obliterate the embedded data. Testing the robustness and security of a watermarking system via attacks is as important as the design process and can be viewed as its inseparable element in a broad sense. A number of attacks as well as some countermeasures have been reported in the literature [7][8]. Most of the previous attacks target at specific types of watermarking schemes, for which analysts have full knowledge of the watermarking algorithms. The analysts are able to perform experiments with many watermarked, non-watermarked, and attacked samples, and to observe the results in real time. In this paper, we will discuss attacks under an emulated rivalry environment in which analysts have no knowledge of the watermarking algorithms. Here we consider attack using blind estimation without priors which is based on the assumption that the watermark is additive.

In this paper we will concentrate on watermarking attack of still images, which is one of major research areas in watermarking technology. Obviously, the attack introduced in the article can be applied to audio and video watermarking algorithms with the safety of generality and technical modifications depending on the physics of the considered media. The paper is organized as follows. The state-of-art in watermark attacks is given in Section 2. A general model of digital watermarking is presented in Section 3. In Section 4, blind source separation problem is reviewed. We consider the watermarked image as linear instantaneous mixtures of the host image, present a blind source separation algorithm. Finally, Section 5 summarizes the main conclusions and discusses the practical implication of this study.

2. State-of-art Watermarking attacks

In watermarking terminology, an attack is any processing that may impair detection of the watermark or communication of the information conveyed by the watermark. The processed watermarked data is then called attacked data. There are two kinds of watermark attacks: non-intentional attacks, such as compression of a legally obtained, watermarked image or video file, and intentional attacks, such as an attempt by a multimedia pirate to destroy the embedded information and prevent tracing of illegal copies of watermarked digital video. Watermarking is treated as a communications problem, in which the owner attempts to communicate over a hostile channel, where the non-intentional and the intentional attacks form the channel. The owner tries to communicate as much watermark information as possible while maintaining a sufficient high data quality. Contrary, an attacker tries to impair watermark communication while impairing the data quality as little as possible. Therefore, digital watermarking scenarios can be considered as a game between the owner and the attacker [9]. Continuing with the analogy of watermarking as a communications system, some researchers have chosen to work on modeling and resisting attacks on the watermark. They work on the philosophy that the more specific the information known about the possible attacks, the better we can design systems to resist it.

In November 1997, the first version of StirMark was published as a generic tool for simple robustness testing of image watermarking algorithms. It introduced random bilinear geometric distortions to de-synchronize watermarking algorithms. In January 1999 the authors discussed the urgent need for fair evaluation procedures for watermarking systems and a first benchmark was made possible with the release of StirMark 3.1. [17][18]

Voloshynovskiy et al. [7] proposed a stochastic formulation of watermarking attacks using an estimation-based concept and a new benchmark. The proposed attacks consist of two main stages: (a) watermark or cover data estimation; (b) modification of watermarked data aiming at disrupting the watermark detection and of resolving copyrights, taking into account the statistics of the embedded watermark and exploiting features of the human visual system. Compared with the model of the Stirmark benchmark, the authors proposed the 6 following categories of tests: denoising attacks and wavelet compression, watermark copy attack, synchronization removal, denoising/compression followed by perceptual remodulation, denoising and random bending. Results indicate that even though some algorithms perform well against the Stirmark benchmark, almost all algorithms perform poorly against their benchmark. This indicates that much work remains to be done before claims about “robust” watermarks can be made.

Su et al. [8] presented a channel model for imperfectly synchronized watermark detection. The focus of their analysis is on blind scalar Costa scheme (SCS) watermarking, which is for perfectly synchronized detection independent from the host signal statistics and thus outperforms the popular spread-spectrum (SS) watermarking by far. Robust watermark detection after desynchronization attacks is still an important problem in the field of digital watermarking. The authors computed the maximum achievable watermark rate for imperfectly synchronized watermark detection within the given channel model. The results show that the characteristics of the host signal play a major role in the performance of imperfectly synchronized watermark detection. Applying these results, the authors proposed a resynchronization method based on a securely embedded pilot signal. The watermark receiver exploits the embedded pilot watermark signal to estimate the transformation of the sampling grid. This estimate is used to invert the desynchronization attack before applying standard SCS watermark detection.

The watermarking game for Gaussian original signals had has been first investigated by Moulin et al. [9], however, with a differently defined attack distortion measure. Moulin et al. derive that the optimum attack under all possible attacks is a specific amplitude scaling and additive white Gaussian noise (SAWGN) attack, the Gaussian test channel (GTC). Eggers et al. [10] translated SAWGN attacks into effective AWGN attacks and presented the capacity analysis of SAWGN attacks. The authors showed that optimal watermark embedding in case of SAWGN attacks produces watermark signals that are correlated with the original signal.

Kundur et al. [11] assert that certain attacks such as cropping, filtering and perceptual coding can be modeled as fading in a noise attenuating non-stationary channel. Thus, the authors employ principles of diversity and channel estimation to improve performance of watermarking schemes. Analysis is provided to show that for common attacks such as spatial cropping and compression, the wavelet-domain, which tends to isolate these distortions, is one of the best domains in which to embed the information. The approach is implemented in a technique known as Robust Reference Watermarking (RRW) that employs watermark repetition and a reference watermark to estimate the attack characteristics. Simulation results verify their observations demonstrating that the class of attacks for which a watermarking scheme is robust can be greatly broadened.

Craver et al. [12] proposed the first protocol attack. They introduce the framework of invertible watermark and show that for copyright protection applications watermarks need to be non-invertible. The idea of inversion consists of the fact that an attacker who has a copy of the watermarked data can claim that the data contains also the attacker's watermark by subtracting his own watermark. This can create a situation of ambiguity with respect to the real ownership of the data. The requirement of non-invertability on the watermarking technology implies that it should not be possible to extract a watermark from non-watermarked image. As a solution to this problem, the authors propose to make watermarks signal-dependent by using a one-way function.

Kutter et al. [13] donated another kind of protocol attack, called watermark copy attack. The concept of the attack consists in copying a watermark from a watermarked image to a target image without using any specific information about the watermarking technology. This new attack has several important implications depending on the application of digital watermarking. If a technology does not resist to the copy attack, a user may not be sure if a detected watermark really belongs to the data under inspection. This is a big problem for many applications. New watermarking technologies should therefore take the watermark copy attack into account during the technology design process.

Previous analytic work in the area of watermark attack has assumed additive Gaussian watermark channels or taken into account the statistical properties of the images and watermarks in the design of attacks. This paper present a general model for watermark attacks based on blind source separation. The new thought is to recover the host image from their linear instantaneous mixtures (watermarked image) without resorting to any prior knowledge.

3. Preliminaries and Notation

We consider digital watermarking a communications problem. Fig. 1 depicts a general perspective on digital watermarking. From now on, variables in bold letters will represent an *M*-dimensional discrete-time/discrete-space process **x**[n] whose elements are independent identically distributed(IID) random variables(RVs)~ $N(0,\sigma_x^2)$. An image **x** is transformed into a watermarked version **s** applying a watermarking embedding function. The inputs of the embedding function also include a secret **K** only known to the copyright owner and a message **m** taken from a finite discrete alphabet with *M* elements. The embedding process is dependent on the key **K** and must be required that the quality difference between x and s is not too large. For embedding, a key sequence *k* of appropriate length is derived from the key **K**. The watermark signal is denoted by the difference **w** = **s**-**x**. The watermarked data **s** might be further processed or even replaced by some other data. This process, denoted attack, produces the attacked data **r**. The attack can be any processing such that the quality difference between x and r is acceptable. Usually, the goal of the attack is to impair or even remove the embedded watermark information. The attacked data **r** is equivalent to the received data **r**, which is input to the watermark reception process. Watermark reception denotes both, decoding of a received watermark message **m'** using key **K** and, watermark detection, meaning the hypothesis test whether **r** is watermarked or not. In some applications of digital watermarking, the original data **x** might be available to the watermark receiver; however, in many applications it is not available. We focus on blind watermarking which denotes the scenario where the watermark receiver operates without access to the original data **x**. Here, **x**, **w**, **s**, **r**, and **k** are an *M*-dimensional discrete-time/discrete-space process, and x_i, w_i, s_i, r_i, and k_i

refer to their respective *n*th elements.

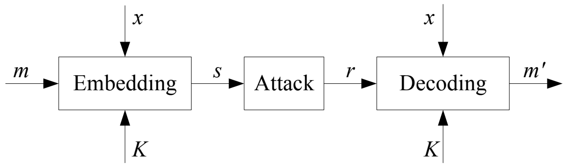

Fig.1 General model of digital watermarking

We treat the original signal **x** as a discrete-time/space and in turn model signals as *M*-dimensional discrete-time/space non-Gaussian random processes. Indexing of an *M*-dimensional signal **x** is denoted by **x**[n], where n=($n_1,\ldots,n_M$). The watermark signal **w** is treated as a discrete-time/space and in turn model signals as ergodic, zero-mean, wide-sense stationary, *M*-dimensional discrete-time/space Gaussian random processes. The watermark signal **w** is represented by the random process **w**[n] and has variance σ_w^2, where n=($n_1,\ldots,n_M$). The original **x**[n] and the watermark **w**[n] are assumed independent of one another.

4. Attack using blind source separation

In blind signal separation (BSS) the goal is to recover the original signals from their mixtures. The only assumption for the methods of BSS is mutual statistical independence of the original signals. Usually, the signals are assumed to be mixed linearly. The data model and basic idea is shown on Fig. 2. Let's denote the source signals (host image) **x**[n] = $[x_1, x_2, \ldots, x_n]^T$, mixed signals(watermarked image) **s**[n] = $[s_1, s_2, \ldots, s_n]^T$ and recovered signals(attacked image) **r**[n] = $[r_1, r_2, \ldots, r_n]^T$, The mixing is static, so the data model can be expressed as follows:

$$\mathbf{s} = A\mathbf{x} \tag{1}$$

where $A \in \Re^{n \times n}$ is an unknown mixing matrix. It's obvious the task can be solved up to scale and permutation. The aim is to find the unmixing matrix B so that:

$$\mathbf{r} = B\mathbf{x} \tag{2}$$

and thus

$$BA = PD \tag{3}$$

where $P \in \Re^{n \times n}$ is a permutation matrix and D = diag(d_1, d_2, …, d_n) diagonal matrix.

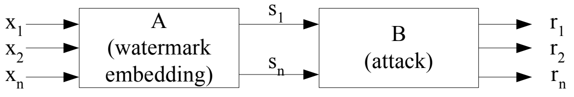

Fig.2 Data model for proposed watermark attack

The application of BSS is in many research areas, such as cocktail party problem, preprocessing for speech recognition tasks, biomedical applications (artifact removal from EEG, extraction of fetal ECG, event-related potentials study and searching for epileptic centra in brain, MRI), GSM communications, cryptography etc [14]. The BSS can be accomplished by various methods. In this paper we discuss the standard Principal Component Analysis (PCA) as a preprocessing step for more sophisticated Independent Component Analysis (ICA) based on higher order statistic (HOS)[15].

BSS task is accomplished by finding demixing matrix B so that the mutual information at the output is I(r_i, r_j) = 0. This is also the basic noise-free definition of ICA (Independent Component Analysis)[16]. It is equivalent to maximization of joint entropy H(r). The maximization of joint entropy (or minimization of mutual information) is performed by introducing higher order statistics (i.e. nonlinearity).

Since we do not know the original pdfs, various functions are used to approximate them. Here we use cumulative density function (cdf), which is defined as bellow:

Let **x** is observed signal, **r** = B**x** is separated signal by B and z is transformed from r by some nonlinearity $z_i = g_i(r_i)$. For a single variable $z = g(r)$, the H(*z*) is maximized when g(·) is a cumulative density function (cdf.) of *r*. Often a standard sigmoid function is commonly used, i.e. $z = (1 + e^{-r})^{-1}$ then f(*r*)=1-2*z* or hyperbolic tangents *z*=tanh(*r*) and f(*r*)=-2*z*. From [15] we have rule:

$$B[k+1] = B[k] + \eta[k]\left(\left[B^T\right]^{-1} + f(r)x^T\right) \tag{4}$$

The stochastic gradient ascent algorithm is easy to implement by means of neural networks. The estimated unmixing matrix B is a matrix of weights of NN. In case of not zero mean sources we should also include bias weights in update but it is better to subtract the mean values from the signals to avoid further updating.

The equation (4) can be implemented in one layer neural network but the convergence will be very slow. So we make as the first step the decorrelation or whitening of the observed data.

$$B[k+1] = B[k] + \eta[k]\left(I - f(r[k])g^T(r[k])\right)B[k] \tag{5}$$

Whitening can be accomplished by PCA method in the first layer of NN simply taking *f*, *g* = *r*. In next layers (at least one) of the designed NN the ICA is simultaneously performed. The design looks like on Fig.3. We can use various nonlinear functions and equation (5) for ICA. Especially for ill-conditioned mixing matrices it's necessary to use more layers with decreasing learning rule. The nonlinearity can be: $f(r) = r^3$, $g(r) = r$; f(r) = r, $g(r)$ = tanh(ar); $f(r) = r^3$, $g(r)$ = tanh(ar).

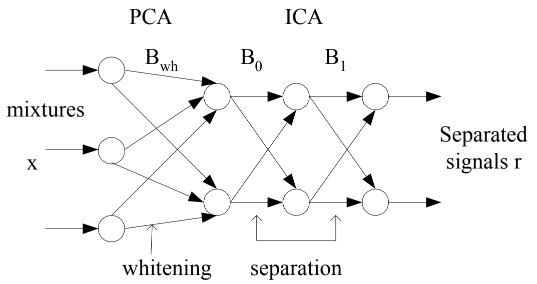

Fig. 3 Multilayer neural network

5. Conclusion and Future Work

A conceptually new watermark attack using blind source separation was analyzed. The basic idea of this kind attack was that source separation algorithms do not assume any a priori information on the watermarking system, the knowledge of watermark embedding, and other physical parameters, which determine the linear system. While much attention has been paid to the blind identification of convolution mixtures, source separation concerns itself only with multiplicative mixtures. In this paper, the algorithm we used is based on the block processing of higher order cumulates. It assumed non Gaussian source signals but do not require the exact knowledge of their distributions. When the source distributions are known in advance, significant improvement can be gained by taking this information into account. In this case, it is possible to implement a maximum likelihood approach to solve the source separation problem. Based on the analysis above, we are going to do the corresponding attacks on commercial watermark schemes to test our models as the basis of a new piracy tool and for further enhancement of the existing watermarking techniques. In the future, we will apply the presented model also for the analysis of geometric signal modifications such as scaling or affine transformations of images and video sequences.

Acknowledgments

This work was supported by the Korea Science and Engineering Foundation (KOSEF) through the Advanced Information Technology Research Center (AITrc)

References

[1] J. R. Hernández, F. Pérez-gonzález, “Statistical analysis of watermarking schemes for copyright protection of images”, Proceeding of the IEEE, Vol. 87, No.7, pp1142-1166, July 1999

[2] R. Bäuml, J. J. Eggers and J. Huber, “ A channel model for desynchronization attacks on watermarks”; Proceedings of SPIE: Electronic Imaging 2002, Security and Watermarking of Multimedia Contents IV, Vol. 4675, San Jose, CA, USA, January 2002.

[3] A. Sequeira, D. Kundur, “Communication and information theory in watermarking: a survey,” Multimedia Systems and Applications IV, Proc. SPIE (vol. 4518), pp. 216-227, Denver, Colorado, August 2001.

[4] S. Voloshynovskiy, A. Herrigel, N. Baumgärtner and T. Pun, “A stochastic approach to content adaptive digital image watermarking”, In International Workshop on Information Hiding, Vol. LNCS 1768 of Lecture Notes in Computer Science, pp. 212-236, Springer Verlag, Dresden, Germany, 29 September -1 October 1999.

[5] J. K. Su, B. Girod, “Fundamental performance limits of power-spectrum condition-compliant watermarks”, Security and Watermarking of Multimedia Contents II, Proc. SPIE (vol. 3971), pp.314-325, San Jose, CA, USA, January 2000.

[6] J.K. Su, J.J. Eggers, B. Girod, “Analysis of digital watermarks subjected to optimum linear filtering and additive noise”, In Signal Processing, Special Issue on Information Theoretic Issues in Digital Watermarking, vol. 81 (6), pp.1141-1175, June, 2001.

[7] S. Voloshynovskiy, S. Pereira, V. Iquise, and T. Pun, “Attack modeling: towards a second generation watermarking benchmark”, In Signal Processing, Special Issue: Information Theoretic Issues in Digital Watermarking, vol. 81 (6), pp. 1177-1214, June 2001.

[8] Jonathan K.Su, Frank Hartung, and Bernd Girod, “A channel model for a watermark attack,” Security and Watermarking of Multimedia Contents, Proc. SPIE Vol.3657, pp159-170, San Jose, California, January 1999.

[9] P. Moulin and J. A. O'Sullivan, “Information-theoretic analysis of information hiding.” Preprint, available at http://www.ifp.uiuc.edu/~moulin/paper.html, 1999.

[10] J.J. Eggers, R. Bäuml, B. Girod, “Digital watermarking facing attacks by amplitude scaling and additive white noise”, 4th Intl. ITG Conference on Source and Channel Coding Berlin, Jan. 28-30, 2002

[11] D. Kundur and D. Hatzinakos, “Diversity and attack characterization for improved robust watermarking,” IEEE Transactions on Signal Processing 29, October 2001.

[12] S. Craver, N. Memon, B. L. Yeo, and M. M. Yeung. Can invisible watermark resolve rightful ownerships? In Fifth Conference on Storage and Retrieval for Image and Video Database, volume 3022, pages 310-321, San Jose, CA, USA, February 1997.

[13] M. Kutter, S. Voloshynovskiy, and A. Herrigel, “Watermark copy attack. ” Security and Watermarking of Multimedia Content II, Proc. SPIE Vol. 3971, pp.23-28 , San Jose, California USA, January 2000.

[14] Zhang Xianda, Bao Zheng, “Communication Signal Processing,” ISBN 7-118-02443-0, 2000.12, in Chinese.

[15] Sejnowski T.J. Bell A.J. “An information-maximization approach to blind separation and blind deconvolution”. Neural Computation, 7:1129-1159, 1995.

[16] A. Hyvärinen. “Survey on independent component analysis. Neural Computing Surveys”, 2:94-128, 1999.

[17] http://www.cl.cam.ac.uk/~fapp2/watermarking/stirmark/index.html

[18] Fabien A. P. Petitcolas, Ross J. Anderson, Markus G. Kuhn. Attacks on copyright marking systems, in David Aucsmith (Ed), Information Hiding, Second International Workshop, IH'98, Portland, Oregon, U.S.A., April 15-17, 1998, Proceedings, LNCS 1525, Springer-Verlag, ISBN 3-540-65386-4, pp. 219-239.