抗可逆性攻击和嵌入器攻击的安全水印系统*

张新鹏, 王朔中
(上海大学通信与信息工程学院, 上海 200072)

摘 要: 提出了可以同时抵抗可逆性攻击和嵌入器攻击的安全水印系统。此系统由协议和算法两个部分组成: 水印协议由权威机构颁布, 规定了合法水印嵌入及检测的框架; 水印算法由嵌入者自行选择或设计。该水印系统将处于不同层次的抗可逆性攻击水印协议与抗嵌入器攻击水印算法结合在一起, 另外借鉴了迭加一次伪随机序列负载多比特信息的思想, 将单比特水印系统扩展为多比特系统。

关键词: 数字水印; 可逆性攻击; 嵌入器攻击
中图分类号: TP391 文献标识码: A 文章编号: 0529-6579(2004)S2-0010-04

数字水印是信息隐藏的重要分支, 用于保护数字多媒体的知识产权[1,2]。但数字水印技术面临着各种各样的攻击[3], 这些攻击大致可以分成四类[4]: 信号处理、几何攻击、协议攻击和工具攻击。信号处理攻击是指对加载了水印信息的信号进行常规处理, 例如压缩、滤波、加噪、重采样等, 这种攻击不一定是恶意的, 但有可能将水印破坏掉。几何失真攻击是指对信号进行缩放、仿射变换、剪切、删除或插入采样点、随机几何失真等处理, 这种攻击破坏了检测同步, 使水印发生畸变无法正确提取。协议攻击可分为拷贝攻击和可逆性攻击, 攻击的是数字水印的概念, 混淆版权归属。拷贝攻击是将数字水印从一份含水印的媒体中复制到另外一份原本不含水印的媒体中。可逆性攻击是使含合法水印的数字媒体中出现另一个水印, 在两个水印同时存在的条件下, 无法确定该媒体的版权归属[5]。工具攻击指攻击者可以利用水印嵌入器或检测器对嵌入的水印信号进行估计, 然后删除水印。

为了抵御这些攻击, 研究者们提出了许多算法、协议来提高数字水印的安全性。事实上真正安全的水印必须能够抵御所有攻击, 因为攻击者可以尝试各种手段直至攻击成功为止。而目前关于数字水印的研究成果大多是面向某个攻击或某类攻击的, 将不同的水印技术结合起来对抗多种攻击还鲜有报导。本文将抗可逆性攻击的水印协议[6]与抗嵌入器攻击的水印算法[7]结合在一起, 提出了可抗这两种攻击的水印系统, 并且借鉴一个伪随机序列负载多比特信息的思想[8], 推广为多比特系统。

1 抗可逆性攻击与抗嵌入器攻击

1.1 抗可逆性攻击的水印协议

可逆性攻击是一种巧妙的攻击方法: 并不试图删除数字媒体中的合法水印, 而是通过逆向构造使数字媒体中出现另外一个水印以混淆版权归属[5]。假设版权所有者 A 用嵌入工具 E_A 将自己的水印 W_A 嵌入原始载体 I 得到合法的含水印拷贝 I', I' 是可以被合法使用者和攻击者得到的。假设攻击者 B 拥有自己的嵌入工具 E_B 和检测工具 D_B, 如果 B 可以伪造出一个水印 W_B 及一份媒体 I_B, 满足下面条件: ① $E_B(I_B, W_B) = I'$; ② $D_B(I', W_B) = 1$(这里检测结果 1 表示"有水印"、0 表示"无水印"); ③ I_B 与 I' 非常相似, 那么水印系统 (E_B, D_B) 称为可溯的。如上操作后, 用 D_A 可以从 I' 中检测出合法水印 W_A, 而用 D_B 可以检测出伪造的水印 W_B, 因此无法判断 I' 的版权归属。也就是说, 如果一个水印系统是可逆的, 就可以被攻击者用来混淆数字产品的版权归属。

为使可逆性攻击不可实现, Craver 建议用不可逆水印算法嵌入版权标识[5], 即所有的合法水印算法必须公开, 并由权威机构检验其不可逆性。当给定了数字媒体 I' 和一个不可逆水印系统 (E_B, D_B) 时, 攻击者是无法伪造水印 W_B 和原始载体 I' 满足上面的 3 个条件。不可逆水印算法不难设计, 令水印信号为原始数据及密钥的单向杂凑输出即可[9]。Katzenbeisser[10] 提出了另外一种抵抗可逆性攻击的

* 收稿日期: 2004-07-01
基金项目: 国家自然科学基金资助项目(60372090); 上海市教委青年基金资助项目
作者简介: 张新鹏(1975 年生), 男, 博士; E-mail: xzhang@staff.shu.edu.cn

方案，即在传统的水印算法上应用一个水印协议，这个水印协议利用了数字签名机制，因此保证了系统的不可逆性，不必再要求水印算法的不可逆性。

本文针对 Craver 等[6]的方案提出了更强的“基于算法构造的可逆性攻击”。就是说，得到含合法水印的数字媒体 I'后，攻击者不仅可以构造 W_B 和 I_B，也可以构造一个不可逆水印算法（E_B，D_B）使得上面的3个条件得到满足。对于 Craver 的不可逆要求，一方面，攻击者可以公开伪造的水印算法，声明算法的不可逆性；另一方面，可以宣称 I'中含有水印 W_B。而对于 Katzenbeisser 的方案，攻击者完全可以在遵从不可逆水印协议的条件下，构造 W_B、I_B、(E_B，D_B) 满足可逆性攻击的3个条件。在这里，不可逆的水印算法(E_B，D_B) 是为媒体 I'特制的，并不一定适用于其它媒体。如果要攻击其它媒体，必须重新构造水印算法。尽管这样，对媒体 I'的攻击已经成功。

在文 [6] 中不仅给出了算法构造可逆性攻击的两个实例，而且提出了一种安全的水印协议，使可逆性攻击与算法构造可逆性攻击均无法实施。这个协议在水印算法公开的条件下（Kerckhoff 原则），将水印检测程序 D 的代码作为一个二进序列 S_D 来处理，并规定了水印算法与嵌入信号之间的联系：

①公开定义单向杂凑函数 h；

②版权所有者 A 选择密钥 k_A 并计算 $W_A = h(S_D, k_A)$，将 W_A 嵌入原始信号生成含水印的载体 $I' = E(I, W_A) = E[I, h(S_D, k_A)]$。

③如果有人宣称拥有对数字媒体的版权，他必须提供原始信号 I，密钥 k 和检测软件 D，当 $D[I', I, h(S_D, k_A)] =$ TRUE 时承认版权归属，否则拒绝。

1.2　抗嵌入器攻击的水印算法

在某些特殊情况下，攻击者可以得到水印嵌入工具甚至密钥，例如在 DVD 版权保护框架里，每个 DVD 刻录器中集成了水印检测器和水印嵌入器，而且所有的检测器和嵌入器共享相同的密钥[11]。此时，攻击者不必知晓水印嵌入和检测的机制，就可以利用嵌入工具擦去已嵌入在数字媒体中的水印。设版权所有者以密钥 K 和嵌入器 E 将水印 W 嵌入原始数字载体 I 得到合法的含水印版本 I'，

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

当攻击者拥有嵌入器 E 和密钥 k 时，可以在 I'上嵌入另一个同样的水印，

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

一般情况下前后两个水印产生的改动是相同的。攻击者得到 I'、I''后，便可构造不含水印的非法拷贝：

$$I_A = 2I' - I'' = I' - (I'' - I') \approx I \tag{3}$$

文献 [3] 给出了一种略加变化的嵌入器攻击方法。

为了抵抗这种攻击，曾提出一种安全的水印算法，其核心思想是：即便嵌入相同的水印，所引起的变化也是互相独立的[7]。嵌入过程如下：

①首先由原始载体数据得到一系列零均值的变换域系数

$$C = \{C(0), C(1), \cdots\cdots, C(N-1)\}$$

②由密钥产生一个置乱函数 f，使得 i 与 $f(i)$ 之间一一对应，其含义是将第 i 个元素置于第 $f(i)$ 个位置。

③随机产生均值为 0、标准差为 σ_w 的独立同分布信号

$$W = \{W(0), W(1), \cdots, W(N-1)\}$$

④将水印信号和一个置乱的版本（置乱方法依赖于密钥）同时被迭加到原始数据上，

$$C''(i) = C(i) + W(i) + W[f^{-1}(i)], \quad i = 0, 1, k, N-1 \tag{4}$$

完成后丢弃水印信号 W。

检测水印是否存在时，不需水印信号，仅有密钥即可。以同样的变换得到 C'后，计算

$$R = \frac{1}{N}\sum_{i=1}^{N} C''[f(i)] \cdot C''(i) \tag{5}$$

参数 R 在 0 附近说明不含水印，R 在 σ_w^2 附近说明含有水印；如果载体受到过嵌入器攻击，R 则会在 $2\sigma_w^2$ 附近，因此嵌入器攻击不但不能删除合法水印，反而因为 R 值的增加留下了攻击的痕迹。

2　安全的单比特水印系统

如前所述，对数字水印技术的攻击多种多样，真正安全的水印系统必须能够有效抵御所有攻击。在文 [6] 和 [7] 中提出的方法可以分别抵抗可逆性攻击和嵌入器攻击，而且处于协议和算法 2 个不同的层次，因此我们可以将这两项工作结合起来，形成一个水印系统，以便同时抵抗可逆性攻击和嵌入器攻击。本节我们考虑单比特水印方案，即水印检测结果仅仅是“有”或“无”两种情况。

安全水印系统由协议和算法两个部分组成。水印协议应当由权威机构颁布，它并不规定具体的水印算法，但是只有遵从该协议进行水印的嵌入和检测才被认为是合法的。水印算法由版权所有者自行设计，当有人宣称拥有某份数字媒体的知识产权时，他必须提供密钥并在权威机构颁布的协议框架下检测到水印才能获得承认。下面我们既给出一种水印协议，又给出一种水印算法，也就是说，当权

威机构颁布了该协议并且版权所有者选用了该算法的条件下，可逆性攻击和嵌入器攻击都无法实施。

2.1 水印协议

①公开定义单向杂凑函数 h。

②设版权所有者 A 选用的水印软件为（E_A，D_A），其密钥为 k_A（密钥空间很大），A 首先计算“派生密钥” $K_A = h(S_D, k_A)$，这里 S_D 是水印提取程序 D_A 的二进代码。然后根据派生密钥 K_A 利用嵌入算法 E_A 将水印嵌入载体数据中，在这里，如何代表 1 *bit* 水印信息并嵌入载体数据由的信号由水印算法和派生密钥 K_A 具体规定。

③如果有人声称拥有某数字媒体的版权，他必须提供密钥 k 和检测软件 D，计算 $K = h(S_D, k)$ 后，由派生密钥 K 和 D 判断该数字媒体是否含有水印。

2.2 水印算法

嵌入算法的输入是原始信号 I、派生密钥 K，嵌入算法如下：

①由原始载体数据得到一系列零均值的变换域系数

$$C = \{C(0), C(1), \cdots\cdots, C(N-1)\}$$

②由派生密钥 K 产生一个置乱函数 f。

③随机产生均值为 0、标准差为 σ_w 的独立同分布信号

$$W = \{W(0), W(1), \cdots\cdots, W(N-1)\}$$

④如式（4）将水印信号和一个用 f 置乱后的版本同时迭加到原始数据上，完成后丢弃水印信号 W。

提取算法如下：

按照协议要求，根据密钥 k 和检测软件 D 得到参考密钥 K 后，产生一个置乱函数 f，如式（5）计算参数 R，并以 $\sigma_w^2/2$ 为阈值，R 小于阈值判为无水印，R 大于阈值判为有水印。

在这个水印系统中，水印协议保证了可逆性攻击无法实施，水印算法保证了嵌入器攻击无法实施。如果攻击者欲进行可逆性攻击或基于算法构造的可逆性攻击，他必须至少找到合适的密钥 k、水印检测软件 D 使得用派生密钥 K 和 D 给出的检测结果为“有水印”。但是水印协议规定了派生密钥 K 是密钥 k 和水印检测软件 D 的单向杂凑结果，所以利用含合法水印的数字媒体逆向构造密钥 k 和水印检测软件 D 是不可能的，其道理类似于文[6] 所述。另一方面，如果攻击者进行嵌入器攻击，水印算法使得每次相同密钥对应的置乱函数是相同的，但每次产生的信号 W 是互相独立的，所以嵌入器攻击只能使 R 值增加，不但不能删除合法水印，反而留下了攻击线索，其道理类似于文[7] 所述。

3 多比特安全水印系统

上节中水印系统的检测结果仅为“有”或“无”，实际上带有具体内容的数字水印更有应用价值。现结合多比特水印技术，将可抗可逆性攻击和嵌入器攻击的单比特水印系统推广为多比特水印系统。

虽然将单比特水印算法反复多次就可以实现多比特的嵌入，但信号失真也增加了多倍。为了在多比特嵌入时保持较好的隐蔽性，可利用一个伪随机序列负载多比特信息，水印提取时比较载体数据与多个伪随机序列的相关值，具有最大相关值的伪随机序列负载对应的多比特信息即嵌入的水印[8]。我们借鉴这样的思想，提出如下的多比特水印系统。

3.1 水印协议

①公开定义单向杂凑函数 h。

②设版权所有者欲嵌入的多比特水印信息为 W，计算

$$W_E = W \| W_1 = W \| h(S_D, k, W)$$

这里，S_D 是水印提取程序 D 的二进代码、k 是密钥，算符“$\|$”表示将两个二进制序列前后衔接起来。然后利用嵌入算法和密钥将 W_E 嵌入载体数据。

③设利用提取程序 D 和密钥 k 从载体数据中提取出的信息为 W_E'，将其分割为 W' 与 W_1'，如果 $W_1' = h(S_D, k, W')$，则判定载体中含有的多比特水印信息为 W'，否则提取水印失败。

3.2 水印算法

嵌入算法的输入是原始信号 I、多比特信息 W_E，嵌入算法如下：

①将多比特信息 W_E 分成 R 段，每段 m bit，m 不宜过大。记这 R 段为 W_1，W_2，……，W_R。

②设用于隐藏信息的非直流系数个数为 N，用密钥 k 产生 $2^m \cdot R$ 种置乱 N 个系数的方法，将这 $2^m \cdot R$ 种置乱方法分成 R 组，每组置乱方法对应于水印信息 W 的某一段。由于每组中含有 2^m 种置乱方法，因此每一种置乱方法又可以对应于这一段水印信息的一种可能（每段水印信息含 m bit，所以有 2^m 种可能）。

③在一组置乱方法中，找到对应那一段的水印信息指示的置乱方法，产生一个零均值随机白信号，将此随机信号及其置乱后的版本迭加在原始系数

上,然后丢弃这个产生出的信号。对每一组置乱方法都如此处理以嵌入一段水印信息,共进行 R 次。

④逆变换得到含水印的媒体数据。

提取算法如下：

①用密钥 k 产生同样的 $2^m \cdot R$ 种置乱方法，并同样分成 R 组。

②对每一组置乱方法，计算载体系数向量与 2^m 个置乱后向量的相关系数，其中最大值对应的置乱方法所指示的 m bit 即嵌入信息的一个分段。

③重复 R 次，将 R 段内容拼接起来即 W_E'。

应用这个水印算法，虽然嵌入了 $m \cdot R$ 比特信息，但实际上仅在原始系数上迭加了 R 个随机信号及其置乱版本，每次迭加都嵌入了 m bit。而且这个算法沿袭了上面两节中的根据载体数据与重排序数据的相关值检测水印的方法，所以嵌入器攻击只能使水印嵌入强度更大，而不会删除水印。如果攻击者进行可逆性攻击或基于算法构造的可逆性攻击，为了水印提取时能够得到想要的结果，攻击者必须找到合适的密钥 k、水印检测软件 D。但是水印协议规定了嵌入的信息 $W_E = W \| h(S_D, k, W)$，也就是说必须嵌入一部分附加信息，这部分信息是密钥 k、水印提取软件 D 和攻击者标识的单向杂凑结果。因此利用含合法水印的数字媒体逆向构造密钥 k 和水印提取软件 D 是不可能的。

4　结　论

本文将处于不同层次的抗可逆性攻击水印协议与抗嵌入器攻击水印算法结合在一起，提出了可同时抵抗这两种攻击的数字水印系统，而且借鉴了迭加一次伪随机序列负载多比特信息的思想，将单比特水印系统扩展为多比特系统。这个初步的水印系统仅仅结合了我们自己的几项有限的工作，距真正可靠的完全安全还有较大的距离，但它是对提高安全性一个有益的尝试，希望这样能够有助于对抗多种攻击，使数字水印技术走向成熟。

参考文献：

[1] PETITCOLAS F A P, ANDERSON R J, KUHN M G. Information hiding? A survey[J]. Proc IEEE, 1999, 87: 1062－1078.

[2] HARTUNG F, KUTTER M. Multimedia watermarking techniques[J]. Proc IEEE, 1999, 87: 1079－1107.

[3] COX I J, LINNARTZ J M G. Some general methods for tampering with watermarks[J]. IEEE J Selected Areas in Communications, 1998, 16: 587－593.

[4] VOLOSHYNOVSKIY S, PEREIRA S, PUN T, et al. Attacks on digital watermarking: classification, estimation-based attacks, and benchmarks[J]. IEEE Communications Magazine, 2001, 118－126.

[5] CRAVER S, MEMON N, YEO B, et al. Resolving rightful ownerships with invisible watermarking techniques: limitations, attacks, and omplications[J]. IEEE J Selected Areas in Communications, 1998, 16: 573－586.

[6] ZHANG X, WANG S. Invertibility attack against watermarking based on forged algorithm and a countermeasure[J]. Pattern Recognition Letters, 2004, 25(8): 967－973.

[7] ZHANG X, WANG S. Watermarking scheme capable of resisting attacks based on availability of inserter[J]. Signal Processing, 2002, 82(11): 1801－1804.

[8] ZHANG X, WANG S, ZHANG K. Multi-bit watermarking scheme based on addition of orthogonal sequences // Computer Network Security[C]. Lecture Notes in Computer Science, 2003, 2776: 407－418.

[9] HWANG M S, CHANG C C, HWANG K F. A watermarking technique based on one-way hash functions[J]. IEEE Trans Consumer Electronics, 1999, 45: 286－294.

[10] KATZENBEISSER S, VEITH H. Securing symmetric watermarking schemes against protocol attacks // Security and Watermarking of Multimedia Contents Ⅳ[C]. Proceedings of SPIE, 2002: 260－268.

[11] BLOOM J A, COX I J, KALKER T, et al. Copy protection for DVD video[J]. Proc IEEE, 1999, 87(7): 1267－1276.

Secure Watermarking System Capable of Resisting Invertibility and Inserter Attacks

ZHANG Xin peng, *WANG Shuo zhong*

(School of Communication and Information Engineering, Shanghai University, Shanghai 200072, China)

Abstract: A secure watermarking system capable of resisting invertibility and inserter attacks is proposed. The proposed watermarking system is made up of two parts: a protocol issued by an authoritative organization, which defines a framework of watermark embedding and detection, and an algorithm chosen or designed by a copyright owner. While the protocol can defeat invertibility attacks, the algorithm ensures that any attack based on availability of inserter is impossible to succeed. By using a previously introduced technique in which each sequence can carry more than one bit without additional distortion, the proposed method can be expanded to a multiple-bit system.

Key words: digital watermarking; invertibility attack; inserter attack