基于可逆信息隐藏技术的认证方案的攻击与改进

王 泓[1], 黄方军[1, 2]

[1] 中山大学 计算机学院, 广东省信息安全技术重点实验室, 广州 中国 510006
[2] 中国科学院信息工程研究所, 信息安全国家重点实验室, 北京 中国 100093

摘要 可逆信息隐藏技术一方面能够对图像的原始性和完整性进行认证和保护, 同时还能够确保无失真地恢复原始图像, 近年来在公安、司法等领域受到越来越多的关注。基于可逆信息隐藏的认证方案需要同时满足可逆和认证两个方面的要求, 在实际中具有较大挑战性, 目前成功案例较少。在文献[1]中, Hong 等人提出了一种新的基于可逆信息隐藏技术的认证方法, 该方法借助 IPVO(Improved pixel-value-ordering)和 LSB(Least significant bit)替换等技术, 可以有效地对图像进行认证。本文我们对 Hong 等人的方法进行了深入研究, 指出在该方法中仅有部分像素参与认证码的生成且攻击方能够很容易地获知这部分像素, 因此在安全性上还存在不足。针对该方法存在的安全漏洞, 我们提出了一种针对性的攻击方案, 即攻击方可选择对图像中未参与认证码生成和嵌入的像素进行修改。该攻击方案可以在不影响所嵌入认证码提取的同时, 实现有意义篡改。为了提高认证算法的安全性, 本文还针对 Hong 等人算法的缺陷提出了相应的改进方案, 即将更多像素引入认证码的生成过程中并在嵌入前对图像块进行置乱。理论分析和实验结果验证了本文提出的攻击和改进方案的有效性。

关键词 可逆信息隐藏; 认证; IPVO; LSB 替换
中图法分类号 TP391 DOI 号 10.19363/J.cnki.cn10-1380/tn.2022.01.04

Attack and Improvement of an Authentication Scheme Based on Reversible Data Hiding

WANG Hong[1], HUANG Fangjun[1, 2]

[1] Guangdong Province Key Laboratory of Information Security Technology, School of Computer Science and Engineering, Sun Yat-Sen University, Guangzhou 510006, China
[2] State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China

Abstract Reversible data hiding (RDH) technology can not only authenticate and protect the originality and integrity of the image, but also ensure that the original image can be restored without distortion. In recent years, it has attracted more and more attention in the fields of public security, justice and so on. Because the image authentication based on reversible data hiding needs to meet the requirements of reversibility and authentication function at the same time, it is quite challenging in practice, and there are few successful cases at present. In reference [1], Hong et al. proposed a new authentication scheme based on reversible data hiding. With the help of improved pixel-value-ordering (IPVO) method and the least significant bit (LSB) substitution technique, the image can be effectively authenticated without losing the reversibility. In this paper, the authentication method proposed by Hong et al. is deeply studied. We point out that in this method only part of pixels participate in the generation of authentication code, and the attackers can easily find them, so there are some deficiencies in security. Aiming at the security vulnerabilities of this method, a targeted attack scheme is proposed, that is, the attackers can choose to modify those pixels which are not involved in the generation of authentication code or the embedding process. Our attack scheme can achieve meaningful tampering without affecting the extraction of the embedded authentication code. Further, in order to improve the security of the reversible authentication algorithm proposed by Hong et al., in this paper we also propose a corresponding improvement scheme according to the defects of Hong's algorithm, that is, more pixels are introduced into the process of generating the authentication code, and besides, the Arnold scrambling is used before embedding the authentication code. Theoretical analysis and experimental results verify the effectiveness of the attack and improvement scheme proposed in this paper.

Key words reversible data hiding; authentication; IPVO; LSB substitution

通讯作者: 黄方军, 博士, 教授, Email: huangfj@mail.sysu.edu.cn。
本课题受国家自然科学基金(No. 62072481, No. 61772572)资助。
收稿日期: 2021-04-02; 修改日期: 2021-08-04; 定稿日期: 2021-11-10

1 引言

近年来, 手机、便携式录音笔、以及摄像机等数字产品的日益普及, 为人们获取图像、音视频等多媒体文件提供了便利。人们可以非常方便地利用这些手持式设备拍照、录音或者摄像, 并将其上传到微信空间、个人微博等。同时, 随着计算机、网络、多媒体处理及传输技术等的飞速发展, 图像抓拍、音视频监控等设备也已遍布日常生活的每个角落。图像和音视频监控技术在社会治理、民生服务等领域得到了长足的发展。但同时我们也应看到, 随着计算机软件技术的迅速发展, 图像、音视频处理软件越来越多, 对图像、音视频等多媒体文件进行篡改也变得越来越容易。如何确保这些海量多媒体文件在存储和传输过程中的原始性和完整性是一个亟待解决的问题。信息隐藏是 20 世纪 90 年代发展起来的一门新的学科。其基本原理是利用人类感官系统对某些细节的不敏感性, 嵌入部分信息到原始载体而不引起观察者的怀疑。面对新形势下的海量多媒体文件, 利用信息隐藏技术来对其进行认证和保护具有非常重要的意义。

然而由于信息嵌入所具有的入侵特性, 信息隐藏在完成信息嵌入的同时会对原始载体引入改变。自 20 世纪 90 年代末期, 着眼于保护军事、医学和艺术等领域的高精度图像, 一门新学科即可逆信息隐藏(又叫可逆水印)受到了国内外学者广泛关注。不同于传统信息隐藏技术, 可逆信息隐藏技术可在提取所嵌入信息的同时无失真恢复原始载体。考虑到原始图像、音视频等文件可能会成为法庭上重要证据, 利用可逆信息隐藏技术对多媒体文件的原始性和完整性进行认证和保护, 同时确保原始载体可恢复, 在实际应用中具有重要意义。

当前应用于认证技术中的可逆信息隐藏主要采用三种方法: 无损压缩(Lossless compression)[2]、差值扩展(Difference expansion, DE)[3]和直方图平移(Histogram shifting, HS)[4]。无损压缩技术通常会在图像视觉质量上引入较大失真, 因此后两者近年来受到更多的关注。差值扩展[3]与直方图平移[4]类似, 都是对载体直方图(如原始直方图和差分直方图等)进行平移和扩展操作来进行信息的嵌入, 其中峰值点被扩展以嵌入秘密信息, 而非峰值点向两侧或某一侧平移以留出嵌入空间。由于平移操作主要用于预留空间, 无信息的嵌入, 通常被称作无效平移。一般图像越平滑, 图像中相邻像素的相关性越高, 产生的载体直方图的峰值越高, 对直方图进行扩展操作得到的嵌入容量越大, 同时无效平移而引入的图像失真越小。为了更好地利用相邻像素的相关性以提高嵌入容量并降低图像失真, 许多学者基于直方图平移与差值扩展技术提出一系列新的基于预测误差直方图的方法, 包括基于一维预测误差直方图的方法[5-6]、基于多维预测误差直方图的方法[7-8]、基于像素值排序的方法[9-15]、基于多直方图的方法[16-17]、基于状态转移矩阵的方法[18-19]、基于非对称直方图[20-21]的方法等。上述方法通常以嵌入容量、视觉质量等作为衡量算法性能的主要指标, 在认证功能方面考虑较少。

由于对图像进行认证需要使认证信息尽可能地在图像中均匀分布, 具体来说是要分布到图像的每个子块, 甚至是每个像素中, 以便于检测出图像被篡改的位置, 因此要在完成认证的同时实现可逆是一项具有挑战性的工作。已有的一些基于信息隐藏的认证方法主要是对图像进行分块认证。如 Lee 和 Suh 利用差值扩展的方法对图像进行认证[22]。认证码的产生利用的是 Holliman 和 Memon 在文献[23]中采取的脆弱水印生成方式, 即首先通过图像自身的信息产生哈希序列, 将哈希序列与待嵌入的二值水印图像异或, 最后将异或后的信息嵌入到图像的每一个子块中。虽然在该方法中哈希序列的生成是由图像信息与密钥共同确定的, 但这种基于块独立的重复嵌入方法易受到伪造攻击[23]。

在文献[24]中, Lo 和 Hu 提出一种新的基于可逆信息隐藏的认证方法。该方法首先将图像分成 N 个子块, 并在每个子块中生成预测误差直方图, 然后对预测误差直方图进行扩展平移操作从而将 N 比特认证信息分别嵌入到 N 个子块中。但考虑到部分子块的嵌入容量可能为 0, 因此对这部分嵌入容量为 0 的子块难以进行有效认证。Nguyen 等人[25]及 Yin 等人[26]分别通过对图像进行小波变换和引入希尔伯特曲线, 对 Lo 和 Hu 的方法进行改进, 从而提高了嵌入容量与认证精度。但二者的方法在安全性上仍存在一定的漏洞, 如可对图像进行以下篡改, 对块中的所有像素值同时加上一个常数(常数攻击)[1], 确保预测误差直方图保持不变, 由于上述的认证方法是通过对预测误差直方图扩展平移来嵌入认证信息的, 因此常数攻击后认证方提取的认证信息与嵌入方嵌入的一致, 篡改无法被检测。

为了进一步提高安全性, Hong 等人[1]提出了基于 IPVO[11]和 LSB 替换的认证方法。在该方法中, Hong 等人根据嵌入容量将子块分为两类, 即嵌入容量大于 0 的子块和嵌入容量等于 0 的子块。该认证

方法不仅能对嵌入容量大于0的子块用IPVO的方法进行认证, 且巧妙地利用了LSB替换对嵌入容量为0的子块进行认证, 有效提高了认证精度。同时对于常数攻击, 该方法也能有效地进行检测。

该方案提出后, 受到国内外学者的较多关注[27-29]。本文针对 Hong 等人提出的方法[1]进行深入研究, 指出该方案中由于攻击方能准确获知认证图像中参与了认证码生成和提取的部分像素, 在安全性方面还存在明显缺陷。针对这些缺陷, 本文提出了一种针对性的攻击方案, 即选择对子块中不参与认证码生成且不影响认证信息提取的像素进行篡改, 从而使得Hong 等人认证方案无法准确检测一些针对性且有意义的篡改。为了提高算法的安全性, 本文还提出了相应的改进方案。理论分析和实验结果验证了本文提出的攻击和改进方案的有效性。

本文的具体安排如下: 第二节简单介绍了IPVO 方法及Hong 等人提出的基于可逆信息隐藏的认证方法, 第三节给出了本文所提出的攻击和改进方案, 并在第四节给出了我们所提出的攻击与改进方案的具体实验结果, 最后在第五节我们对本文进行了总结。

2 相关工作

Hong 等人提出的基于可逆信息隐藏的认证方法主要基于 IPVO 方法来实现, 这一部分我们将首先介绍 IPVO 方法, 再对 Hong 等人提出的认证方法进行分析。

2.1 IPVO 方法[11]

IPVO 是对 Li 等人提出的 PVO 方法[9]的改进, 两者均采用分块嵌入方案。不失一般性, 设图像中任意子块像素为$\{x_1, x_2, \ldots, x_{n-1}, x_n\}$。首先对子块中的像素按像素值的大小升序排序, 得到排序后的像素值序列为$\{x_{\sigma_1}, x_{\sigma_2}, \ldots, x_{\sigma_{n-1}}, x_{\sigma_n}\}$, 其中下标$\sigma_i$表示$x_{\sigma_i}$在初始序列$\{x_1, x_2, \ldots, x_{n-1}, x_n\}$中的位置。排序方式为稳定排序, 即当$x_i = x_j$且$i < j$时, 有$\sigma_i < \sigma_j$。PVO 方法直接采用最大像素值减第二大像素值, 最小像素值减第二小像素值来得到预测误差值, 并进而构造预测误差直方图。通过改变每个子块内的最大和最小像素以实现对预测误差直方图的扩展和平移操作, 进而完成信息的嵌入。PVO 方法中, 一般选择对预测误差值 1 和–1 进行扩展, 对小于–1 和大于 1 的预测误差值分别向左右两边平移以腾出空间。

为了进一步提高嵌入容量, IPVO 在计算预测误差值时考虑了像素值的位置信息, 从而可以利用预测误差值 0 进行扩展嵌入。预测误差值的计算如公式(1)和(2)所示

$$\begin{aligned} d_{\max} &= x_u - x_v \\ u &= \min(\sigma_n, \sigma_{n-1}) \\ v &= \max(\sigma_n, \sigma_{n-1}) \end{aligned} \tag{1}$$

$$\begin{aligned} d_{\min} &= x_s - x_t \\ s &= \min(\sigma_1, \sigma_2) \\ t &= \max(\sigma_1, \sigma_2) \end{aligned} \tag{2}$$

在上述公式中, 当$\sigma_n < \sigma_{n-1}$时, $u = \sigma_n$, $v = \sigma_{n-1}$, 由排序的方式可知此时有$d_{\max} > 0$; 当$\sigma_n > \sigma_{n-1}$时, $u = \sigma_{n-1}$, $v = \sigma_n$, 根据排序的方式可知此时有$d_{\max} \leqslant 0$。对于$d_{\min}$同理。

分别构造关于$d_{\max}$与$d_{\min}$的预测误差直方图, 具体信息嵌入如公式(3)和(4)所示

$$x'_{\sigma_n} = \begin{cases} x_{\sigma_n} + b, & d_{\max} = 0 \ or \ d_{\max} = 1 \\ x_{\sigma_n} + 1, & d_{\max} < 0 \ or \ d_{\max} > 1 \end{cases} \tag{3}$$

$$x'_{\sigma_1} = \begin{cases} x_{\sigma_1} - b, & d_{\min} = 0 \ or \ d_{\min} = 1 \\ x_{\sigma_1} - 1, & d_{\min} < 0 \ or \ d_{\min} > 1 \end{cases} \tag{4}$$

在公式(3)和(4)中, b表示要嵌入的认证信息比特。由公式(3)可见: 当嵌入信息为 b = 1 时, 该子块内的最大像素值x_{σ_n}加 1, 此时预测误差值如为 0 则扩展为 –1, 如为1则扩展为2; 当嵌入信息为b = 0时, 该子块内的最大像素值保持不变, 此时预测误差值无论是0还是1均保持不变; 当预测误差值小于0或大于 1 时, 对预测误差直方图进行平移操作, 即对最大像素值加 1, 使得小于 0 和大于 1 的预测误差值分别向左右两边平移。公式(4)对该子块内最小像素值的操作同理可得。由嵌入规则可知, 嵌入信息后任意子块中最大像素值x'_{σ_n}和最小像素值x'_{σ_1}仍为该子块中的最大和最小值。

对于提取方, 同样对子块中的像素按像素值大小升序排序, 排序后用同样的方法计算预测误差, 根据预测误差值提取信息并恢复图像。在提取公式(5)和(6)中, b'为提取的认证信息比特。由公式(5)和(6)可知: 当预测误差值为 0 或 1 时, 提取认证信息比特$b' = 0$; 当预测误差值为-1或2时, 提取认证信息比特$b' = 1$。

$$b' = \begin{cases} 0, & d_{\max} = 0 \ or \ d_{\max} = 1 \\ 1, & d_{\max} = -1 \ or \ d_{\max} = 2 \end{cases} \tag{5}$$

$$b' = \begin{cases} 0, & d_{\min} = 0 \ or \ d_{\min} = 1 \\ 1, & d_{\min} = -1 \ or \ d_{\min} = 2 \end{cases} \tag{6}$$

原始载体图像按照公式(7)和(8)进行恢复。

$$x_{\sigma_n} = \begin{cases} x'_{\sigma_n}, & d_{\max} = 0 \ or \ d_{\max} = 1 \\ x'_{\sigma_n} - 1, & d_{\max} < 0 \ or \ d_{\max} > 1 \end{cases} \tag{7}$$

$$x_{\sigma_1}=\begin{cases}x'_{\sigma_1}, & d_{\min}=0\ or\ d_{\min}=1\\ x'_{\sigma_1}+1, & d_{\min}<0\ or\ d_{\min}>1\end{cases} \tag{8}$$

2.2 Hong 等的的认证方法[1]

在 Hong 等人的方法中, 首先对图像分块, 并按嵌入容量将块分成可嵌入块和非可嵌入块, 然后针对不同类别的块分别用 IPVO 和 LSB 替换方式进行认证。具体的认证方案如下。

2.2.1 块分类

在Hong等人的基于可逆信息隐藏的认证方法中, 首先对原始载体图像进行4×4不重叠分块, 再将所得的每一个4×4图像块进一步划分成 4 个2×2的子块, 具体划分方式如图 1 所示。为了便于说明, 我们将4×4的图像块称为**嵌入块**, 将2×2的子块称为**嵌入单元**, 根据嵌入块的嵌入容量又将其分为可嵌入块(即 E 块)和非可嵌入块(即 U 块), 其中 E 和 U 分别代表 Embeddable 和 Unembeddable, 其具体分类方案如下:

1) 对每个2×2嵌入单元中的像素按像素值大小升序排序;

2) 计算2×2嵌入单元中的预测误差值$d_{\max}$与$d_{\min}$, 根据其为 0 和 1 的数量计算该嵌入单元的嵌入容量, 由 IPVO 算法可知每个2×2嵌入单元的嵌入容量最多为 2;

3) 将 4 个2×2嵌入单元的嵌入容量相加得到整个嵌入块的嵌入容量, 整个嵌入块的嵌入容量最多为 8, 最少为 0;

4) 当整个嵌入块的嵌入容量为 0 时将其分类为 U 块, 否则分类为 E 块。

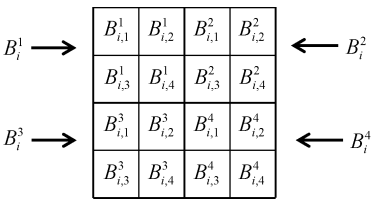

图 1 块B_i的划分

Figure 1 The partition of block B_i

2.2.2 U 块认证

首先对 U 块进行预处理, 具体如下: 1) 对所有2×2嵌入单元所属的预测误差直方图进行平移操作, 如当原始嵌入单元中得到的预测误差值为–1 或 2 时, 平移后分别为–2 或 3; 2) 在平移后的嵌入块中根据密钥 *key* 任意选择一个嵌入单元, 将该嵌入单元中的最小像素值与最大像素值进一步向左右两边各平移两个单位, 即进行减 2 和加 2 的操作。上述预处理主要用于确保在该嵌入块的任意2×2嵌入单元中, 所得的预测误差值均一直大于 2 或小于–1, 从而使认证方在提取信息过程中不至于与 E 块混淆。

对预处理后 U 块的认证方案如下: 1) 根据当前嵌入块的所有像素 7 个高位平面的值及当前嵌入块的位置信息生成 1 比特的认证码; 2) 用认证码替换根据秘钥 *key* 选定的2×2嵌入单元中左上角像素的 LSB 位, 对该嵌入块进行认证, 并将被替换的 LSB 顺次保存在数组$S=[s_1,s_2,\cdots,s_n]$中。

2.2.3 E 块认证

对E块进行认证无需进行预处理, 记当前E块的嵌入容量为α。对 E 块认证过程具体如下: 1) 对每个2×2嵌入单元中的像素按其像素值大小升序排序, 分别提取每个嵌入单元中居中的 2 个像素, 共得到 8 个像素; 2) 根据这 8 个像素的像素值及当前嵌入块的位置信息生成$\alpha-1$比特的认证码, 记当前的认证码为 A(注: 当嵌入容量$\alpha=1$时, 直接生成 1 比特认证码即可); 3) 从 2.2.2 节中得到的 S 数组中顺次提取 1 比特, 并将其和 A 进行拼接, 得到的α比特记为A', 即$A'=A\,|\,s_i$(注: 当$\alpha=1$时, 认证码不与s_i进行拼接, 即$A'=A$); 4) 将α比特的A'用 IPVO 方法嵌入到当前嵌入块中。

3 攻击与改进方案

根据前文介绍可知, Hong 等人的方法中, 由于分块信息是公开的, 攻击方无需其他信息即可准确计算出每一个嵌入块的嵌入容量, 再根据嵌入容量可将其准确分类为 U 块或 E 块。同时对于嵌入容量不为 0 的 E 块, 攻击方也可以根据 IPVO 的嵌入和提取规则, 准确得到嵌入方具体通过改变哪些像素值嵌入信息。根据上述分析, 我们提出如下的攻击及改进方案。

3.1 攻击方案

攻击方按照已公开的分块信息, 对图像进行分块并将嵌入块划分成 4 个嵌入单元。计算出嵌入块的嵌入容量后进行块分类, 若为 E 块则按照一定规则修改嵌入单元的最大或最小像素, 若为 U 块则按照一定规则修改嵌入单元中的非左上角像素。攻击方案的流程如图 2 所示。

3.1.1 E 块攻击

认证方在对 E 块进行认证时, 首先提取出所嵌入的认证信息, 然后根据原嵌入方所采取的方式重新生成认证信息, 将提取出的认证信息与重新生成

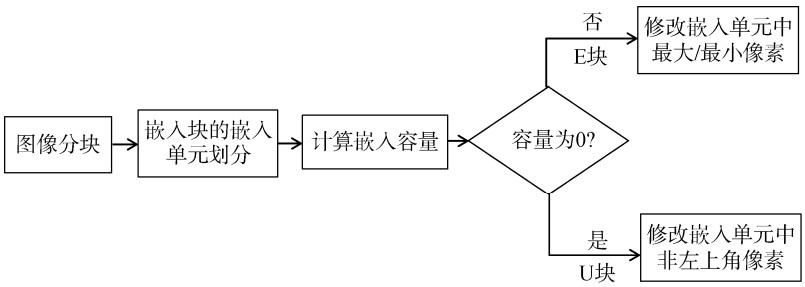

图 2　攻击方案流程图

Figure 2　The flowchart of the attack

的认证信息进行比对, 若二者一致, 则表明该E块没有被篡改, 反之则表明该 E 块已被篡改。因此对 E 块攻击后要达到的主要目标包括: 1) E 块篡改前后认证方提取的认证码保持不变; 2) E 块的篡改不影响认证信息的生成, 即篡改前后运用原嵌入方所采取的方式生成的认证信息保持不变。

根据公式(5)和(6)IPVO 算法的提取规则, 只有当预测误差值属于{–1, 0, 1, 2}时, 认证方才可提取出对应的认证信息, 因此要保证 E 块篡改前后认证方提取的认证码保持不变, 只需在篡改前后确保与预测误差值{–1, 0, 1, 2}相关的所有像素值保持不变即可。另一方面, 由于E块认证码的生成只与E块中每个2×2嵌入单元中的两个中间像素相关。因此, 要使得对 E 块的篡改不影响认证信息的生成, 只需确保每个2×2嵌入单元的中间两个像素保持不变即可。

由上述的分析可知, 篡改方可对不参与认证码生成且不影响认证信息提取的像素进行篡改。具体到每一个2×2嵌入单元, 只要在该嵌入单元中计算得到的预测误差值不属于$\{-1, 0, 1, 2\}$, 则可对最大或最小像素值进行修改。具体篡改方案如下:

1) 对E块中2×2嵌入单元的像素按其值的大小升序排序, 由公式(1)与(2)分别计算出预测误差值$d_{\max}$与$d_{\min}$;

2) 当$d_{\max}\notin\{-1,0,1,2\}$时, 将$2\times2$嵌入单元最大像素值增加为不大于 255 的任意值;

3) 当$d_{\min}\notin\{-1,0,1,2\}$时, 将$2\times2$嵌入单元最小像素值减小为不小于 0 的任意值。

经过篡改, E 块中嵌入单元的最大和最小像素值仍为嵌入单元中最大和最小值, 不会对 E 块的嵌入容量产生影响, 也就不会影响认证方对嵌入块的分类。而篡改仅发生在$d_{\max}\notin\{-1,0,1,2\}$与$d_{\min}\notin\{-1,0,1,2\}$的那些嵌入单元的最大和最小像素值上, 对认证信息的提取与重新生成均不会产生影响。

3.1.2　U 块攻击

认证方在对 U 块进行认证时, 首先提取出所嵌入的认证信息, 然后根据原嵌入方所采取的方式重新生成认证信息, 将提取出的认证信息与重新生成的认证信息进行比对, 若两者一致, 则表明该U块没有被篡改, 反之则表明该 U 块已被篡改。

根据上文介绍, Hong 等人的方法中, 对于 U 块嵌入方仅利用预处理后图像的当前嵌入块所有像素的 7 个高位平面来生成认证码, 因此只需保证当前嵌入块所有像素的 7 个高位平面不变, 即可确保篡改前后生成的认证码保持不变。

此外, 由该认证方法可知, 嵌入方用密钥 *key* 随机选择嵌入块中任意一个2×2嵌入单元, 并用 1 比特认证码替换该2×2嵌入单元左上角像素的 LSB。虽然篡改方无法准确得知嵌入方所选中的2×2嵌入单元, 但篡改方只需保证每个2×2嵌入单元的左上角像素保持不变即可, 从而可确保篡改前后从载体图像中提取的认证码保持不变。

根据上述分析, 在对U块进行攻击时, 首先要确保每个2×2嵌入单元的左上角像素保持不变。另外, 由于U块所有像素的高7位均用于认证码的生成, 所以对 U 块的攻击只考虑对每个2×2嵌入单元的非左上角像素 LSB 位进行翻转攻击。但根据 2.2.3 节可知, U 块是非可嵌入块, 在其每一个2×2嵌入单元中均有$d_{\max}\notin\{-1,0,1,2\}$与$d_{\min}\notin\{-1,0,1,2\}$成立, 因此若随机对非左上角像素的 LSB 位进行翻转攻击, 有可能导致$d_{\max}\in\{-1,0,1,2\}$或$d_{\min}\in\{-1,0,1,2\}$, 从而进一步导致认证方无法准确定位 U 块和 E 块。因此对 U 块的篡改可采取如下方式:

1) 计算每个2×2嵌入单元的预测误差值$d_{\max}$与$d_{\min}$;

2) 在确保$d_{\max}\notin\{-1,0,1,2\}$和$d_{\min}\notin\{-1,0,1,2\}$的情况下, 对每一个$2\times2$嵌入单元的非左上角像素 LSB 位进行随机翻转。

3.2 改进算法

由于对 E 块像素值的修改幅度更大, 对像素的篡改效果更加明显, 第 4 节中将给出实验对此进行说明, 因此对于 E 块考虑更加安全的认证码生成策略。

根据上文提出的攻击方案可知, 由于在 Hong 等人的认证方法中可嵌入块 E 块的每个2×2嵌入单元中的最大和最小像素值均不参与认证码的生成, 因此可对其进行较大幅度的修改。为了提高认证算法的安全性, 可以考虑将 E 块中的所有像素均引入到认证码的生成过程。具体实施过程中, 可以考虑在嵌入认证信息前, 根据该 E 块的所有 16 个像素及其位置信息来生成认证码, 并将其按 Hong 等人的方法嵌入到相关嵌入块中。不同之处在于, 认证方在进行认证时, 先按公式(7)和(8)对原始载体进行恢复, 并根据恢复后得到的嵌入块重新生成认证码, 然后将其与提取的认证信息进行比对以确定当前可嵌入块是否被篡改。

图 3 说明了改进算法与 Hong 等人的认证算法对 E 块认证时的区别, 图 3(a)和(b)中的红色区域分别是 Hong 等人的算法与改进算法中参与认证码生成的像素, 与 Hong 等人的认证算法相比, 改进算法在生成认证码时考虑了嵌入单元中的最大和最小像素值。

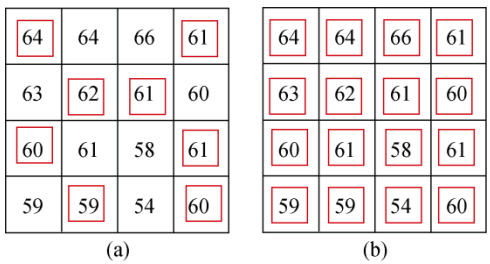

图 3 Hong 等人算法(a)与改进算法(b)参与认证码生成的像素对比

Figure 3 The difference of the pixels participating in the generation of the authentication code between the Hong's algorithm (a) and the improved algorithm (b)

此外, 根据 Hong 等人的算法可知, 分块信息是公开的, 因此攻击方能够根据嵌入块的划分方案, 计算得到嵌入块的嵌入容量, 进而对其进行准确分类。无论对于 E 块还是 U 块, 攻击方能够准确得知嵌入方针对每个2×2嵌入单元的预测误差直方图所采取的扩展或平移等操作。为了避免攻击方获知嵌入块的划分方案, 可在划分嵌入单元之前对嵌入块进行 Arnold 置乱, 再将置乱后的嵌入块进一步划分成 4 个2×2嵌入单元。在 Arnold 置乱的过程中需要 3 个正整数参数, 分别记为 a, b, n, 其中 a, b 为映射矩阵的参数, n 为迭代次数, 三个参数数值过大时易导致计算速度降低, 本文将 a, b, n 的大小均控制在[1, 256]之间, 则每个参数均用 8 比特即可存储。用密钥 *key* 对置乱参数进行加密, 将其以差值扩展的方式嵌入已嵌入认证信息的图像中。认证方同样以差值扩展的方式提取并解密得到置乱参数。

至此, 改进算法可总结如下:

1) 对原始载体图像进行4×4不重叠分块, 对每个4×4嵌入块进行 Arnold 置乱, 将置乱后的嵌入块按图 1 的方式进一步划分成 4 个2×2嵌入单元, 并将置乱参数存于数组 S 中;

2) 同样根据嵌入块的嵌入容量将其分类为 U 块和 E 块;

3) 对于 U 块, 按与 Hong 等人的方法相同的方式进行认证码的生成与嵌入;

4) 对于 E 块, 根据所有像素的像素值及当前嵌入块的位置信息生成认证码, 认证码的长度与嵌入容量α的对应关系与 Hong 等人方法中的相同, 然后同样用 IPVO 方法将α比特信息嵌入到嵌入块中。

改进算法与 Hong 等人的方法相比, 将可嵌入块 E 块中更多的像素引入了认证码的生成过程, 同时在对所有嵌入块划分嵌入单元之前增加了对嵌入块置乱的过程。由于置乱后嵌入块中相邻像素的相关性减弱, 一定程度上会使嵌入容量降低, 但基本不会对认证产生影响, 第 4 节中将会有相关实验结果对此进行进一步的说明。

4 实验结果

实验过程中我们主要针对图 4 中的 6 副512×512的灰度图像进行了测试, 具体包括攻击实验和改进算法实验两个部分。攻击实验部分我们主

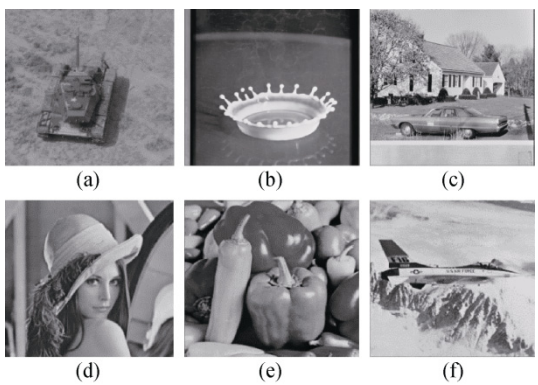

图 4 实验图像 (a) Tank (b) Splash (c) House (d) Lena (e) Pepper (f) Plane

Figure 4 Experimental images (a) Tank (b) Splash (c) House (d) Lena (e) Pepper (f) Plane

要通过篡改像素个数及篡改前后图像之间的峰值信噪比(Peak signal-to-noise rate, PSNR)来对篡改效果进行说明。在改进算法实验部分, 我们从 PSNR、随机攻击的检测率及用本文攻击方法针对性攻击后改进算法的检测效果三个方面来对本文提出的改进算法的效果进行说明。

4.1 攻击实验

攻击实验包括对 E 块和对 U 块的攻击两个方面, 我们运用前文所给出的攻击方案分别进行了实验, 具体如下。

4.1.1 E 块攻击实验

首先用 Hong 等人的方法对图 4 所示 6 幅灰度图像进行认证信息的嵌入, 再用本文提出的攻击方法对图像中的 E 块进行攻击。为使篡改效果更加显著, 当2×2嵌入单元中的最大像素值可修改时, 将其均修改为 255; 当2×2嵌入单元中的最小像素值可修改时, 将其均修改为 0(注: 根据上文的 E 块攻击策略, 上述篡改不影响认证方认证信息的生成和提取, 即认证方无法确认该图像已被篡改)。具体实验结果如表 1 所示。

表 1 E 块篡改实验结果

Table 1 Results of tampering E blocks

图像	E 块个数	可篡改像素个数	PSNR
Tank	13517	79650	11.50
Splash	15631	71830	11.21
House	13381	60362	12.13
Lena	14061	73900	11.37
Pepper	13594	79182	10.90
Plane	14591	64122	11.69
Average Value	14129	71508	11.50

表 1 中的第一列为本文实验中所采用的所有图像, 第二列是每幅图像中 E 块的个数(注: 整幅图像所包含4×4分块的个数为 16384), 第三列是图像可篡改像素个数, 第四列是篡改前和篡改后图像之间的 PSNR 值。由表 1 可见, 在所有 6 幅图像中 E 块比例均较高(E 块个数最少的 House 图像, 其 E 块比例高于 81%), 可被篡改的像素个数也相对较多, 篡改后的图像质量有了大幅度的下降, 表中最小的 PSNR 值为 10.90, 最大的 PSNR 值小于 12.20, 平均值仅为 11.50。说明对 E 块的攻击可以取得非常显著的篡改效果。

图 5 给出了对运用 Hong 等人方法嵌入认证信息后 Lena 图像中所有的 E 块进行针对性攻击后的结果, 可以看到被篡改后的图像会呈现类似于添加了椒盐噪声的效果, 这是因为篡改时部分像素值被修改为 255, 部分像素值被修改为 0, 且可篡改的像素在图像中分布相对均匀。实验过程中我们还对图 5 用 Hong 等人的方法进行认证, 经过验证, 从图 5 中提取的认证信息与按照 Hong 等人方法生成的认证信息完全匹配, 即 Hong 等人所提出的认证方法无法检测出图 5 中的篡改。

图 5 对所有 E 块攻击后的 Lena 图像

Figure 5 Lena image after attacking all E blocks

考虑到图 5 中的篡改类似于添加随机噪声, 在图 6 中我们给出了有意义篡改后的图像, 即对使用 Hong 等人方法嵌入认证信息后的图像中部分 E 块, 运用本文 E 块攻击方案进行篡改后, 使得图像中出现图 6(c)所示的中国银行图标。图 6(a)是对嵌入认证信息后的 Lena 图像在其左上角进行局部篡改后的图像, 篡改后图像左上角处显现图 6(c)所示的中国银行的图标。图 6(b)为对嵌入认证信息后的 Lena 图像在多个位置进行有意义篡改后的图像, 篡改后图像

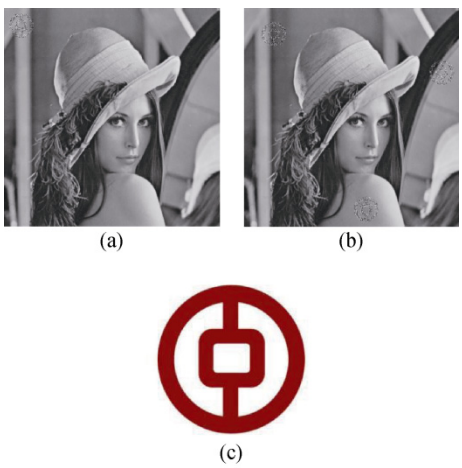

图 6 部分 E 块攻击后图像 (a)左上角篡改后 Lena 图像 (b)多个位置篡改后 Lena 图像(c)中国银行图标

Figure 6 The Lena image after tampering partial E blocks (a) The Lena image after tampering in the upper left corner (b) The Lena image after tampering in multiple locations (c) The bank of China icon

中的左上角处, 镜子处, 以及人物肩膀下方均出现图 6(c) 所示中国银行图标。经过验证, Hong 等人的认证方法无法检测出图 6(a)和(b)中的篡改。

4.1.2 U 块攻击实验

同样, 首先对图 4 中的 6 幅灰度图像用 Hong 等人的方法进行认证信息嵌入, 再用本文提出的攻击方法对 U 块进行攻击。具体实验结果如表 2 所示。

表 2 U 块篡改实验结果

Table 2 Results after tampering U blocks

图像	U 块个数	可篡改像素个数	PSNR
Tank	2867	26084	58.15
Splash	753	5676	64.70
House	3003	29847	57.56
Lena	2323	21540	58.98
Pepper	2790	24792	58.37
Plane	1793	17027	60.00
Average Value	2255	20824	59.63

表 2 第一列为本文实验中所采用的所有图像, 第二列是每幅图像中的 U 块个数(注: 整幅图像所包含4×4分块的个数为 16384), 第三列是图像被篡改的像素个数, 第四列是篡改前和篡改后图像之间的 PSNR 值。由表 2 可见, 每幅图像 U 块的个数, 通常要远低于嵌入容量大于 0 的 E 块的个数, 分别在 750~3800 不等, 但平均篡改的像素个数仍有 20000 左右。虽然本文对 U 块的攻击仅针对像素的 LSB 位进行翻转(即对像素进行±1操作), 篡改前后图像之间的 PSNR 值平均约为 59.63, 通常不能获得有意义的篡改图像, 但从图像真实性和完整性认证的角度而言, 本文所提出的 U 块攻击方案是有效的。

4.2 改进实验

根据第 3.3 节中的改进方案可知, 改进算法需要对4×4嵌入块(包括 E 块和 U 块)中的像素进行 Arnold 置乱。置乱操作通常会减小相邻像素之间的相关性, 导致后续得到的预测误差直方图峰值降低, 因此在认证信息嵌入过程中会增加无效平移的次数, 影响后续所得到的认证图像的视觉效果。但由于 Hong 等人算法中所选择的嵌入块一般为4×4大小, 因此置乱对图像像素之间相关性的影响相对较小。

表 3 分别给出了运用 Hong 等人算法嵌入认证信息后得到的图像, 以及运用本文所提出的改进方案嵌入认证信息后得到的图像与原始载体之间的 PSNR 值, 可以看到运用本文的改进算法后 PSNR 值的下降平均不超过 0.3%。

表 3 PSNR 对比

Table 3 Comparison of PSNR

图像	Hong 等人的算法	改进算法
Tank	50.24	49.90
Splash	49.78	49.38
House	50.44	50.01
Lena	50.66	50.07
Pepper	50.29	49.98
Plane	50.20	50.75
Average Value	50.27	50.01

表 4 随机攻击后正确检测率对比

Table 4 Comparison of correct detection rates of the random attack

图像	Hong 等人的算法(%)	改进算法(%)
Tank	95.39	95.10
Splash	98.10	97.60
House	96.62	96.62
Lena	96.93	95.70
Pepper	97.23	97.23
Plane	95.70	96.01
Average Value	96.66	96.38

表 4 对比了针对随机攻击的检测效果。随机攻击具体方式为: 随机选择图像中 1%左右的嵌入块, 并将所选中块里面的每个像素的像素值随机修改为>0 且<255 的任意值。运用 Hong 等人的方法和本文改进算法对篡改后图像进行认证的正确检测率如表 4 所示, 其中正确检测率为正确检测出被篡改的嵌入块个数与真正被篡改的嵌入块个数的比值。由表 4 可知, 本文所提出的改进方案不影响 Hong 等人算法在普通随机攻击下的检测效果。注: 在 Hong 等人的方法中, 因为图像可嵌入块中可嵌入的认证码数量一般较少, 且为比特串, 根据篡改后图像生成的认证码有一定几率与从篡改后图像中提取的认证码正好相同, 因此通常正确检测率小于 100%。

图 7 给出了改进方案对本文所提出的针对性攻击的篡改检测效果。首先用改进算法在 Lena 图像中嵌入认证信息, 然后运用本文提出的攻击方法对图像中部分嵌入块进行针对性的篡改攻击, 使得篡改后图像中平滑程度较高的左上角处、镜子处和平滑程度较低的帽子与头发交界处出现如图 6(c)的中国银行图标, 得到的篡改后的图像如图 7(a)所示。对图 7(a)运用改进方法进行篡改检测, 将检测到被篡改的嵌入块中所有像素值置为 255, 得到检测结果如图 7(b)所示。图 7(b)中白色区域(像素值为 255 的区域)

为运用本文改进方案检测并定位到的篡改区域, 可以看到改进算法能检测出针对性攻击并对篡改进行定位, 且对于不同平滑程度的区域均有成效。可见改进算法对本文提出的针对性攻击方案有较好的防御作用, 较大程度上提高了 Hong 等人算法的安全性。

(a) (b)

图 7 改进算法对针对性攻击的篡改检测结果(a)针对性攻击后 Lena 图像(b)篡改检测结果

Figure 7 The tampering detection results of targeted attacks for the improved algorithm (a) Lena image after the targeted attack (b) The tampering detection result

5 结语

基于可逆信息隐藏的认证技术能够在对图像的原始性和完整性进行认证和保护的基础上, 实现原始载体图像的无失真恢复。本文主要对 Hong 等人提出的一种新型认证方法进行深入研究, 主要贡献包括: 1) 对 Hong 等人所提出的认证方案的安全性进行了深入分析和研究, 指出了其在安全性方面存在的缺陷, 并提出了一种针对性的攻击策略; 2) 对 Hong 等人的方案进行了改进, 提高了 Hong 等人方案的安全性; 3) 本文所提出的攻击和改进策略对于所有基于 PVO 系列的认证方案具有一定的参考价值, 对后期设计更安全的认证方案具有较为重要的意义。

参考文献

[1] Hong W, Chen M J, Chen T S. An Efficient Reversible Image Authentication Method Using Improved PVO and LSB Substitution Techniques[J]. *Signal Processing*: *Image Communication*, 2017, 58: 111-122.

[2] Barton J M. Method and Apparatus for Embedding Authentication Information within Digital Data. U.S.: 5646997 [P], 1997-07-08.

[3] Tian J. Reversible Data Embedding Using a Difference Expansion[J]. *IEEE Transactions on Circuits and Systems for Video Technology*, 2003, 13(8): 890-896.

[4] Ni Z C, Shi Y Q, Ansari N, et al. Reversible Data Hiding[J]. *IEEE Transactions on Circuits and Systems for Video Technology*, 2006, 16(3): 354-362.

[5] Tsai P, Hu Y C, Yeh H L. Reversible Image Hiding Scheme Using Predictive Coding and Histogram Shifting[J]. *Signal Processing*, 2009, 89(6): 1129-1143.

[6] Li X L, Yang B, Zeng T Y. Efficient Reversible Watermarking Based on Adaptive Prediction-Error Expansion and Pixel Selection[J]. *IEEE Transactions on Image Processing*, 2011, 20(12): 3524-3533.

[7] Xiao M Y, Li X L, Wang Y Y, et al. Reversible Data Hiding Based on Pairwise Embedding and Optimal Expansion Path[J]. *Signal Processing*, 2019, 158: 210-218.

[8] Ou B, Li X L, Zhang W M, et al. Improving Pairwise PEE via Hybrid-Dimensional Histogram Generation and Adaptive Mapping Selection[J]. *IEEE Transactions on Circuits and Systems for Video Technology*, 2019, 29(7): 2176-2190.

[9] Li X L, Li J, Li B, et al. High-Fidelity Reversible Data Hiding Scheme Based on Pixel-Value-Ordering and Prediction-Error Expansion[J]. *Signal Processing*, 2013, 93(1): 198-205.

[10] Wang D W, Zhang X Q, Yu C Q, et al. Reversible Data Hiding by Using Adaptive Pixel Value Prediction and Adaptive Embedding Bin Selection[J]. *IEEE Signal Processing Letters*, 2019, 26(11): 1713-1717.

[11] Peng F, Li X L, Yang B. Improved PVO-Based Reversible Data Hiding[J]. *Digital Signal Processing*, 2014, 25: 255-265.

[12] Ou B, Li X L, Zhao Y, et al. Reversible Data Hiding Using Invariant Pixel-Value-Ordering and Prediction-Error Expansion[J]. *Signal Processing*: *Image Communication*, 2014, 29(7): 760-772.

[13] Ou B, Li X L, Wang J W. High-Fidelity Reversible Data Hiding Based on Pixel-Value-Ordering and Pairwise Prediction-Error Expansion[J]. *Journal of Visual Communication and Image Representation*, 2016, 39: 12-23.

[14] He W G, Zhou K, Cai J, et al. Reversible Data Hiding Using Multi-Pass Pixel Value Ordering and Prediction-Error Expansion[J]. *Journal of Visual Communication and Image Representation*, 2017, 49: 351-360.

[15] Zhang T, Li X L, Qi W F, et al. Location-Based PVO and Adaptive Pairwise Modification for Efficient Reversible Data Hiding[J]. *IEEE Transactions on Information Forensics and Security*, 2020, 15: 2306-2319.

[16] Li X L, Zhang W M, Gui X L, et al. Efficient Reversible Data Hiding Based on Multiple Histograms Modification[J]. *IEEE Transactions on Information Forensics and Security*, 2015, 10(9): 2016-2027.

[17] Qin J Q, Huang F J. Reversible Data Hiding Based on Multiple Two-Dimensional Histograms Modification[J]. *IEEE Signal Processing Letters*, 2019, 26(6): 843-847.

[18] Zhang X P. Reversible Data Hiding with Optimal Value Transfer[J]. *IEEE Transactions on Multimedia*, 2013, 15(2): 316-325.

[19] Zhang W M, Hu X C, Li X L, et al. Optimal Transition Probability of Reversible Data Hiding for General Distortion Metrics and Its Applications[J]. *IEEE Transactions on Image Processing*, 2015, 24(1): 294-304.

[20] Chen X Y, Sun X M, Sun H Y, et al. Histogram Shifting Based Reversible Data Hiding Method Using Directed-Prediction Scheme[J]. *Multimedia Tools and Applications*, 2015, 74(15): 5747-5765.

[21] Kim S, Qu X C, Sachnev V, et al. Skewed Histogram Shifting for Reversible Data Hiding Using a Pair of Extreme Predictions[J].

IEEE Transactions on Circuits and Systems for Video Technology, 2019, 29(11): 3236-3246.

[22] Lee S K, Suh Y H, Ho Y S. Reversible Image Authentication Based on Watermarking[C]. *2006 IEEE International Conference on Multimedia and Expo*, 2006: 1321-1324.

[23] Holliman M, Memon N. Counterfeiting Attacks on Oblivious Block-Wise Independent Invisible Watermarking Schemes[J]. *IEEE Transactions on Image Processing*, 2000, 9(3): 432-441.

[24] Nguyen T S, Chang C C, Yang X Q. A Reversible Image Authentication Scheme Based on Fragile Watermarking In Discrete Wavelet Transform Domain[J]. *AEU - International Journal of Electronics and Communications*, 2016, 70(8): 1055-1061.

[25] Yin Z X, Niu X J, Zhou Z L, et al. Improved Reversible Image Authentication Scheme[J]. *Cognitive Computation*, 2016, 8(5): 890-899.

[26] Nguyen T S, Vo P H. Reversible Image Authentication Scheme Based on Prediction Error Expansion[J]. *Indonesian Journal of Electrical Engineering and Computer Science*, 2021, 21(1): 253.

[27] Bolourian Haghighi B, Taherinia A H, Mohajerzadeh A H. TRLG: Fragile Blind Quad Watermarking for Image Tamper Detection and Recovery by Providing Compact Digests with Optimized Quality Using LWT and GA[J]. *Information Sciences*, 2019, 486: 204-230.

[28] Hurrah N N, Parah S A, Sheikh J A. Embedding In Medical Images: An Efficient Scheme for Authentication and Tamper Localization[J]. *Multimedia Tools and Applications*, 2020, 79(29/30): 21441-21470.

[29] Zhou X Y, Hong W, Weng S W, et al. Reversible and Recoverable Authentication Method for Demosaiced Images Using Adaptive Coding Technique[J]. *Journal of Information Security and Applications*, 2020, 55: 102629.

王泓 于 2019 年在中山大学计算机科学与技术专业获得学士学位。2021 年在中山大学计算机技术专业获得专业硕士学位。研究领域为可逆信息隐藏。研究兴趣: 可逆认证、可逆信息隐藏。Email: wangh265@mail2.sysu.edu.cn

黄方军 于 2005 年在华中科技大学获得博士学位。现任中山大学计算机学院教授、博士生导师。研究领域为多媒体信息隐藏与数字取证、AI 安全-对抗样本攻击与防御。研究兴趣: 多媒体信息隐藏与数字取证、AI 安全-对抗样本攻击与防御。Email: huangfj@mail.sysu.edu.cn