一种具有高攻击类型判别能力的图像空域半脆弱水印算法

肖 磊
(温州大学物理与电子信息工程学院 温州 325035)

摘 要 在考虑图像对比度特性的基础上,提出了一种图像空域半脆弱水印算法。算法由图像对比度敏感性确定各像素水印嵌入比特位,通过LSB(Least Significant Bit)替换方法自适应嵌入水印,从理论上推导出用于图像认证的篡改检测阈值。对图像对比度特性的充分考虑,可确保算法具有较好的透明性。实验表明,算法对可接受的偶然攻击操作具有一定的鲁棒性,同时对恶意攻击较为脆弱,并且能准确定位图像篡改区域。此外,算法能正确区分偶然攻击与恶意攻击,显示出比同类算法更好的攻击类型判别能力。

关键词 半脆弱水印,对比度敏感性,图像认证,篡改检测
中图法分类号 TP309.2,TP391 **文献标识码** A

Image Spatial Semi-fragile Watermarking Algorithm with High Classification Capability of Attack Types

XIAO Lei
(College of Physics & Electronic Information Engineering, Wenzhou University, Wenzhou 325035, China)

Abstract An image spatial semi-fragile watermarking algorithm was proposed by exploiting the contrast sensitivity of the host image. The watermark bit was embedded into the host image by adaptive least significant bit (LSB) substitution and the bit-plane of each pixel for data embedding was determined by the contrast sensitivity. The algorithm deduced theoretically the adaptive threshold for automatic tampering detection and location, and has good transparency because of the exploiting of the contrast sensitivity. Experimental results show that the proposed scheme has good robustness against admissible incidental attacks, while it is sensitive to malicious attacks, and even can localize the tampered region precisely. In addition, this algorithm can distinguish effectively incidental attacks from malicious ones and outperforms other semi-fragile watermarking algorithms at the classification of attacks.

Keywords Semi-fragile watermarking, Contrast sensitivity, Image authentication, Tamper detection

1 引言

随着互联网的迅速普及,人们开始越来越广泛地以数字媒体的形式进行相互交流与通信,这极大地方便了人们的生活;同时多媒体信息处理技术的迅猛发展,也使得攻击者可以较容易地分发、存储、复制,甚至随意篡改未授权的数字作品。然而在用于刑事侦察的法庭证据照片、电子商务领域的电子票据、医学图像和军事遥感图像等要求真实性与完整性的应用中[1,2],一些通常的无意操作(低通滤波、图像压缩等)或有意的、恶意的攻击(如裁剪、涂改等)均可能会造成媒体数据发生改变,进而影响多媒体数据通常的使用与认证。这就需要对多媒体数据进行认证,近年来半脆弱水印技术作为一种有效的多媒体内容认证手段受到了信息安全研究人员的广泛关注。半脆弱水印要求对恶意攻击敏感,同时对一些内容保持的常规信号处理操作(偶然攻击)是鲁棒的[1,3],因而优良的半脆弱水印技术应能有效地区分多媒体数据所遭受的攻击类型。文献[4]利用混沌映射对初值的敏感性与JPEG压缩过程中DCT系数的不变特性,提出了一种抗JPEG有损压缩的半脆弱图像数字水印算法。该算法可检测与定位恶意篡改,同时对JPEG有损压缩具有良好的鲁棒性,但未能考虑究竟如何区分恶意篡改与偶然攻击。Xiao等利用拉普拉斯锐化理论设计了一种半脆弱水印算法[5],该算法能较好地抵抗拉普拉斯锐化操作,但对平均滤波和中值滤波敏感。丁文霞、卢焕章等[6]以彩色图像亮度分量为载体,利用LSB替换方法实现了一种彩色图像空域半脆弱水印算法,该算法具有良好的透明性与安全性,对裁剪、任意涂改、"椒盐"噪声等处理具有一定的鲁棒性,而对高斯噪声、乘性噪声、滤波、亮度变化、对比度改变、平滑、锐化、JPEG压缩、旋转等攻击操作是脆弱的,未能较好区分图像所遭受的攻击类型。Zhao等[7]结合人类视觉系统屏蔽特性提出了一种小波域半脆弱水印算法,算法对JPEG压缩有一定鲁棒性,但对轻微图像滤波和噪声迭加是脆弱的。Woo等提出的一种可自我恢复的半脆弱水印算法[8],通过对载体图像下采样获得水印信号,然后将其嵌入到小波域水平与垂直细节子带,算法对高质量JPEG压缩、轻微局部失真与噪声污染是鲁棒的,但对平均滤波是脆弱的。王向阳等提出了一种基于内容的彩色图像半脆弱水印算法[9],其将小波近似系数进行混沌调制以生成基于图像内容的水印信号,然后结合视觉感知特性及局部系数的相关特性,

到稿日期:2009-10-16 返修日期:2009-12-01 本文受国家自然科学基金项目(60970065)资助。
肖 磊 男,硕士,讲师,主要研究方向为多媒体技术、信息安全、数字水印,E-mail:shidiancn@yahoo.com.cn。

通过分块量化措施将水印信号嵌入到载体图像小波域中。算法具有较好的透明性，对压缩、叠加噪声、平滑滤波等常规图像处理操作有较好的鲁棒性，且能够对剪切、替换等恶意图像篡改做出报警并确定被篡改位置，然而该算法篡改检测阈值的选取为人工凭经验确定，缺乏足够的理论依据。实际上，我们知道，轻微的噪声迭加、低通滤波只对载体图像质量有较小的影响，不应将其判定为恶意攻击。为能实现有效判别图像攻击类型，本文拟结合图像对比度特性，利用自适应 LSB 替换方法，提出一种图像空域半脆弱水印算法。

2 结合图像对比度特性的半脆弱水印算法

2.1 水印嵌入

为获得较好的攻击类型判别能力，本文在充分考虑图像对比度特性的基础上，利用 LSB 替换方法实现了一种图像空域半脆弱水印算法，使各像素用于水印嵌入的比特位与图像内容相自适应。水印嵌入具体步骤如下。

步骤 1　提取原载体图像 I 的高 h 个 MSB(Most Significant Bit)位(如 $h=3$)比特信息，得到残差图像 I_h。

步骤 2　由 Weber 定律有：人眼对不同灰度具有不同的敏感性，通常对中等灰度最为敏感，而对低灰度和高灰度的敏感性则呈非线性下降趋势[10]。于是，可利用残差图像 I_h 由下式计算出各像素 $I(i,j)$的水印嵌入比特位 $\lambda(i,j)$：

$$\lambda(i,j)=round\left(\frac{(7-h)\times|x_h(i,j)-127.5|}{127.5}\right) \tag{1}$$

其中，$x_h(i,j)$为残差图像 I_h 像素 $I_h(i,j)$所对应的像素值，$round(\cdot)$表示四舍五入运算。

步骤 3　读取一幅大小为 $m\times n$ 的二值水印图像为水印，记为 W，由密钥 key 生成二值伪随机序列 $PN=\{pn(i,j)\mid pn(i,j)\in\{0,1\},0\leqslant i\leqslant m-1,0\leqslant j\leqslant n-1\}$。为增强算法安全性，使水印免受攻击者的非法使用或篡改，将水印 W 与伪随机序列 PN 进行异或以获得待嵌入水印信号 W_1。

$$W_1=\{w_1(i,j)\mid w_1(i,j)=w(i,j)\otimes pn(i,j),0\leqslant i\leqslant m-1,0\leqslant j\leqslant n-1\} \tag{2}$$

其中，$\otimes$表示异或运算。

步骤 4　应用 LSB 替换方法以水印比特 $W_1(i,j)$替换像素 $I(i,j)$的第 $\lambda(i,j)$位比特。

步骤 5　重复步骤 4 直到全部水印信息 W_1 嵌入完毕，即得到含水印图像 I_w。

2.2 水印提取

水印提取过程与嵌入过程类似，其详细过程描述如下。

步骤 1　提取待认证图像 I'的高 h 个 MSB 位，获得残差图像 I_h'。

步骤 2　利用式(1)由残差图像 I_h'计算出各像素水印嵌入比特位 $\lambda(i,j)$。实际上，由式(1)可知 $0\leqslant\lambda(i,j)\leqslant 7-h$，即水印的嵌入在各像素 LSBs 位进行，而未改变图像像素高 h 个 MSB 位比特，故水印嵌入前后计算出的各像素水印嵌入比特位 $\lambda(i,j)$是一致的。

步骤 3　提取待测图像 I'各像素 $I'(i,j)$的第 $\lambda(i,j)$位比特得到水印信号 $W_1'=\{w_1'(i,j)\mid w_1'(i,j)\in\{0,1\}\}$，$i=0,1,\cdots,m-1$，$j=0,1,\cdots,n-1$。

步骤 4　由密钥 key(与水印嵌入阶段相同)生成二值伪随机序列 PN，利用 PN 对水印信号 W_1'进行混沌解调制，即恢复出水印 W'，$W'=\{w'(i,j)\mid w'(i,j)\in\{0,1\},0\leqslant i\leqslant m-1,0\leqslant j\leqslant n-1\}$。

2.3 图像认证

由式(3)计算出水印差图像 Q。

$$Q=\{q(i,j)\mid q(i,j)=|w'(i,j)-w(i,j)|,0\leqslant i\leqslant m-1,0\leqslant j\leqslant n-1\} \tag{3}$$

水印差图像 Q 中白色区域(像素值为 1)表示检测错误区域，而黑色区域(像素值为 0)对应水印提取正确的位置。实际上，偶然攻击的影响在差图像中的表现为一些孤立点，这时提取错误的水印比特分散地分布在差图像或提取水印图像中；而遭受恶意攻击后，错误检测点在水印差图像中较为集中。在差图像中，如果一个错误检测点，其所在 8 邻域中至少还存在一个错误检测点，则认为该错误检测点为稠密点；若其所在 8 邻域中没有一个错误检测点，则可认为是稀疏点。不妨先做以下定义。

$\phi_s=\{$水印差图像中稀疏点个数$\}$

$\phi_d=\{$水印差图像中稠密点个数$\}$

$\phi_e=\{$水印差图像中错误检测点个数$\}$

$\phi=\{$水印差图像中像素点个数$\}$

$$\rho=\frac{\phi_s}{\phi}=\frac{\phi_s}{m\times n},\mu=\frac{\phi_d}{E(\phi_e)}$$

其中 $\omega E(\cdot)$表示求数学期望。

据此定义如下的图像认证规则。

(1)如果 $\rho=0$，则图像未被修改；

(2)如果 $\rho>0$，且 $\mu<T$，则认为是偶然攻击，其中 T 是预先确定的阈值，$T\in(0.5,1)$；

(3)如果 $\rho>0$，且 $\mu\geqslant T$，则认为图像遭受到恶意攻击。

对于遭受到恶意攻击的待测图像，在篡改检测图中将稠密点对应的图像区域置为白色像素，即可得到一个较为紧凑的篡改区域，从而实现篡改定位。

2.4 篡改检测阈值的设定

由于经混沌调制后的水印信号 W_1 趋近于均匀分布，各水印比特 $W_1(i,j)$取值 0 或 1 的概率均为 0.5，因此水印差图像中各像素点取值为 0 或 1 的概率也均为 0.5。令 e_{ij}，s_{ij}，d_{ij} 分别为水印差图像中错误检测点、稀疏点和稠密点，$P_{e_{ij}}$，$P_{s_{ij}}$，$P_{d_{ij}}$ 分别是 e_{ij}，s_{ij}，d_{ij} 的出现概率，于是有

$$\begin{cases}P_{e_{ij}}=\dfrac{1}{2}\\ P_{s_{ij}}=\dfrac{1}{2}\left(\dfrac{1}{2}\right)^8=\dfrac{1}{512}\\ P_{d_{ij}}=\dfrac{1}{2}\left(1-\left(\dfrac{1}{2}\right)^8\right)=\dfrac{255}{512}\end{cases} \tag{4}$$

显然 e_{ij}，s_{ij}，d_{ij} 服从(0-1)分布，于是可导出其各自概率、数学期望和方差。

$$\begin{cases}E(e_{ij})=P_{e_{ij}}=\dfrac{1}{2}\\ E(s_{ij})=P_{s_{ij}}=\dfrac{1}{512}\\ E(d_{ij})=P_{d_{ij}}=\dfrac{255}{512}\end{cases} \tag{5}$$

$$\begin{cases}D(e_{ij})=P_{e_{ij}}(1-P_{e_{ij}})=\dfrac{1}{4}\\ D(s_{ij})=P_{s_{ij}}(1-P_{s_{ij}})=\dfrac{1}{512}\times\dfrac{511}{512}\\ D(d_{ij})=P_{d_{ij}}(1-P_{d_{ij}})=\dfrac{255}{512}\times\dfrac{257}{512}\end{cases} \tag{6}$$

根据德莫佛-拉普拉斯定理，当 $m\times n$ 较大时，可近似认为 ϕ_e，ϕ_d 均服从正态分布，于是有

$$\begin{cases} E(\phi_e)=E(\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}e_{ij})=m\times n\times E(e_{ij})=\dfrac{m\times n}{2} \\ E(\phi_d)=E(\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}d_{ij})=m\times n\times E(d_{ij})=\dfrac{255}{512}\times m\times n \end{cases} \tag{7}$$

$$\begin{cases} D(\phi_e)=D(\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}e_{ij})=m\times n\times D(e_{ij})=\dfrac{m\times n}{4} \\ D(\phi_d)=D(\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}d_{ij})=m\times n\times D(d_{ij})=\dfrac{255}{512}\times\dfrac{257}{512}\times m\times n \end{cases} \tag{8}$$

给定虚警率 10^{-9}，则可由式(9)求解阈值。

$$P(\phi_d\geqslant\xi)=1-P(\phi_d<\xi)\approx 1-\Phi\left(\frac{\xi-E(\phi_d)}{\sqrt{D(\phi_d)}}\right)$$

$$=1-\Phi\left(\frac{\xi-0.4980\times m\times n}{\sqrt{0.2500\times m\times n}}\right) \tag{9}$$

求得阈值 $\xi=0.4890\times m\times n-5.9978\times\sqrt{0.2500\times m\times n}$，当 $m=n=256$ 时，可求得篡改检测阈值 T 为

$$T=\frac{\xi}{E(\phi_e)}=\frac{0.4890\times m\times n-5.9978\times\sqrt{0.2500\times m\times n}}{m\times n/2}$$

$$=0.9546 \tag{10}$$

由此可认为：当 $\rho>0$ 时，如果 $\mu>T$，则图像内容被恶意篡改，反之则认为待测图像遭受到偶然攻击。

3 实验结果与性能分析

仿真实验中选取若干以大小为 256×256 的灰度图像为原始载体图像进行了测试，水印图像为 256×256 的二值图像，测试时选取载体图像的高 3 个(即 $h=3$)MSB 位参与计算各像素水印嵌入比特位。以 Barbara 图像为例，图 1 给出了一个水印嵌入示例，可以发现含水印图像视觉效果较好，人眼察觉不出含水印图像中水印信息的存在。

(a)原图像 (b)二值水印图像 (c)含水印图像(37.02dB) (d)提取水印

图 1 水印嵌入示例

3.1 透明性实验

本文采用峰值信噪比 PNSR 来评价含水印图像的质量，实验中针对各种不同类型的灰度图像进行了透明性测试，表 1 列出了一些常见的测试图像的实验结果，从中可知，针对不同类型的图像，该算法所产生的含水印图像 PSNR 值平均高达 37.16 dB 以上，说明算法具有较好的水印透明性。

表 1 含水印图像 PSNR 值

测试图像	含水印图像 PSNR 值(dB)
F-16	38.43
Boat	36.47
Lochness	36.98
Houses	34.20
Baboon	37.76
Barbara	37.02
Opera	39.36
Lena	38.32
Pills	37.82
Goldhill	36.22
Peppers	36.15
平均	37.16

3.2 攻击类型判别实验与性能比较

给定阈值 $T(T=0.9546)$，应用 2.3 节的图像认证规则，本文针对一些通常的攻击操作测试了算法判别攻击类型的能力，相关实验结果如表 2 所列。算法合理地实现了攻击类型的自动分类，有效地将轻微噪声迭加、低通滤波与 JPEG 压缩判别为偶然攻击，而一些半脆弱水印算法[5-8]却将这些操作划分为恶意攻击。然而轻微噪声迭加、低通滤波与 JPEG 压缩等操作仅仅对图像造成小的视觉影响，不应被定为恶意攻击。这表明本文算法具有更好的攻击分类性能。

表 2 攻击类型判别性能比较(其中×表示恶意攻击，√表示偶然攻击)

攻击	ρ	μ	攻击类型				
			本文算法	Xiao 等方法[5]	文献[6]方法	Zhao 等方法[7]	Woo 等方法[8]
JPEG 压缩(100)	0.0723	0.0035	√	√	×	√	√
JPEG 压缩(90)	0.1216	0.0102	√	×	×	√	√
JPEG 压缩(80)	0.2673	0.3204	√	×	×	×	×
10%椒盐噪声	0.1490	0.4983	√	×	√	×	×
1%高斯噪声	0.3131	0.7561	√	×	×	×	×
3×3 中值滤波	0.1037	0.1464	√	×	×	×	×
3×3 均值滤波	0.3268	0.6547	√	×	×	×	×
拉普拉斯锐化	0.1340	0.5641	√	√	×	×	×
裁剪 16×16 区域	0.0005	0.9848	×	×	√	×	×
旋转 1°	0.4590	0.9783	×	×	×	×	×
剪切粘贴 1/4 区域	0.1273	0.9786	×	×	√	×	×
裁剪 1/4 区域	0.1469	0.9859	×	×	√	×	×

3.3 对偶然攻击的鲁棒性

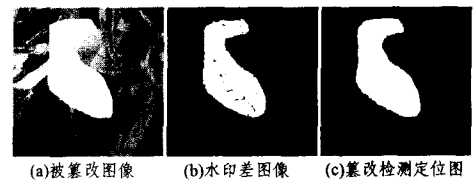
表 3 给出了算法对通常的内容保持攻击操作的鲁棒性测试，其提取水印均具有较高的归一化相关值(NC 值)，特别对中值滤波、轻微噪声迭加、JPEG 压缩等操作的鲁棒性较高，这表明算法对偶然攻击具有较好的鲁棒性。

表 3 鲁棒性实验结果

攻击操作	NC 值
质量因子为 100 的 JPEG 压缩	0.9277
质量因子为 90 的 JPEG 压缩	0.8784
3×3 中值滤波	0.8963
拉普拉斯锐化	0.8660
10%椒盐噪声	0.8510
1%高斯噪声	0.6869

3.4 篡改检测性能

图 2 以剪切攻击实验为例，测试了算法篡改检测性能，从图中可知，本文方案对恶意篡改攻击非常敏感，并能对相应的篡改区域进行较为准确的定位。



(a)被篡改图像 (b)水印差图像 (c)篡改检测定位图

图 2 篡改检测实验结果

结束语 本文在结合图像对比度特性的基础上，提出一种图像自适应的空域半脆弱水印算法。嵌入水印比特深度自

适应于载体图像特征，算法具有较好的透明性。实验表明算法对通常的内容保持操作具有较好的鲁棒性，同时对恶意攻击是脆弱的，且能有效定位恶意篡改区域。此外，算法从概率统计角度分析确定篡改检测阈值，从而实现了攻击类型的自动判别，并能合理区分恶意攻击和偶然攻击。与其他半脆弱水印算法相比，该算法具有更为合理有效的攻击类型判别能力。

参 考 文 献

[1] 杨义先，钮心忻. 数字水印理论与技术[M]. 北京：高等教育出版社，2006

[2] 胡玉平，陈志刚. 用于图像认证的小波域半易损水印算法[J]. 电子学报，2006，34(4)：653-657

[3] 李春，黄继武. 一种抗 JPEG 压缩的半脆弱图像水印算法[J]. 软件学报，2006，17(2)：315-324

[4] 李赵红，侯建军. 基于 JPEG 不变量和混沌映射的半脆弱水印技术[J]. 计算机工程与应用，2007，43(32)：40-43

[5] Xiao J，Wang Y. A semi-fragile watermarking tolerant of Laplacian sharpening[C]// Proc. 2008 International Conference on Computer Science and Software Engineering. Wuhan，China，IEEE Computer Society，2008：579-582

[6] 丁文霞，卢焕章，王浩，等. 一种基于混沌的彩色图像空域半脆弱水印算法[J]. 国防科技大学学报，2008，30(4)：59-63，102

[7] Zhao Y，Sun X H. A Semi-fragile watermarking algorithm based on HVS model and DWT[C]// Proc. 2008 International Conference on Computer Science and Software Engineering. Wuhan，China，IEEE Computer Society，2008：638-641

[8] Woo C S，Jiang D，Binh P. Semi fragile watermark with self authentication and self recovery[J]. Malaysian Journal of Computer Science，2009，22：64-84

[9] 王向阳，杨红颖，侯丽敏. 一种新的半脆弱彩色图像数字水印算法[J]. 自动化学报，2007，33(6)：561-566

[10] 沈兰荪，张菁，李晓光. 图像检索与压缩域处理技术的研究[M]. 北京：人民邮电出版社，2008

[11] 莫佳，谭素琴. 一种基于矩阵分解的脆弱水印算法[J]. 重庆工学院学报：自然科学版，2009，23(5)：107-111

全国第 16 届计算机辅助设计与图形学(CAD/CG'2010)学术会议

2010 年 7 月 28 日－30 日　　中国·太原

http://tiger.cs.tsinghua.edu.cn/cadcg2010　cadcg2010@gmail.com

由中国计算机学会主办，清华大学、太原理工大学联合承办的全国第 16 届计算机辅助设计与图形学学术会议(CAD/CG'2010)将于 2010 年 7 月 28 日在中国太原举行。本次会议内容包括大会学术报告、计算机辅助设计与图形学热点问题专题研讨、最新成果和应用系统演示，并将邀请国内外学术界和产业界的著名专家到会作特邀报告。

会议优秀论文推荐至《计算机辅助设计与图形学学报》(EI 核心)、《工程图学学报》、《系统仿真学报》、《计算机科学》、《中国图象图形学报》等中文核心刊物。大会录用论文将正式结集，由清华大学出版社出版。热诚欢迎一切从事计算机辅助设计与图形学研究、应用及软件开发的专家、学者和专业技术人员参会。

会议论文主题包括(但不限于)：

1. 图形学：图形学基础理论与算法，真实感图形，非真实感图形，工程图形及应用，计算机图形仿真。

2. 可视化、图像与视频处理：科学计算可视化，图形图像融合技术，图像情感计算，计算机动画，虚拟现实与混合现实。

3. CAD/CAM/CAE：计算机辅助设计(CAD)，计算机集成制造，虚拟设计与制造，数字媒体技术与数字内容处理，网络化制造，人机交互技术。

4. EDA 及 VLSI 设计与测试：电子设计自动化(EDA)，VLSI 系统设计方法，VLSI 测试。

5. 几何造型与处理：计算机辅助几何设计，几何造型与处理。

6. 其他与计算机辅助设计与图形学相关的领域。

主办单位：中国计算机学会　　**承办单位**：清华大学、太原理工大学

截稿日期：2010 年 3 月 31 日　　**投稿邮箱**：cadcg2010@cs.tsinghua.edu.cn，cadcg2010@gmail.com

电话/传真：010-62773440　　**联 系 人**：赵康　博士