适用于双重压缩环境的图像内容认证水印算法

付剑晶^{1,2*}, 陈德人¹, 徐达文³, 毛家发⁴

1. 浙江大学计算机科学与技术学院, 杭州 310058

2. 浙江传媒学院新媒体学院, 杭州 310018

3. 宁波工程学院电子与信息工程学院, 宁波 315016

4. 浙江工业大学计算机科学与技术学院, 杭州 310023

* 通信作者. E-mail: fjjsn@163.com

收稿日期: 2018-07-28; 接受日期: 2018-12-17; 网络出版日期: 2019-04-11

浙江省自然科学基金 (批准号: LY18F020003) 和国家自然科学基金 (批准号: 61771270, 61502415) 资助项目

摘要 数字图像内容的恶意篡改可能导致严重后果. 随着图像分辨率的提高及其在网络上流通数量的快速增长, 图像往往要经过双重压缩: 压缩 (JPEG/JPEG2000) 发布 → 解码 → 应用领域的处理 (常规信号处理/安全性攻击/内容的恶意篡改) → 再次压缩 (JPEG2000/JPEG) 发布, 才能进入末端检测. 如何在双重压缩环境下有效判定图像内容是否遭篡改及定位篡改位置是个亟待解决的问题. 本文提出了一种基于旋转向量的新颖的水印表示方法及其调制算法, 并在此基础上建立了半脆弱水印方案用于图像内容认证. 理论分析了水印的稳定性, 使该水印方案具有理论基础; 详细阐述了以特征抽取与重构、水印嵌入与提取、篡改检测与定位为主要内容的认证方案; 分析了水印的鲁棒性、安全性和相关检测性能. 理论分析与实验表明: 该方案水印透明性好, 针对不同的攻击鲁棒性好且分布稳定; 在双重压缩环境下能有效区分恶意篡改与保持内容的处理并定位篡改区域; 安全性好, 能抵制水印攻击、拼贴攻击和伪造攻击. 与相关方案相比, 本文方案综合性能优越, 适用于双重压缩环境下的内容认证, 扩大了基于水印的内容认证的应用范围.

关键词 图像水印, 半脆弱水印, 恶意篡改, 内容认证, 双重压缩

1 引言

图像内容认证用于判定图像内容的真实性, 并可能实现篡改定位以判定篡改动机, 因而认证算法必须对保持图像内容处理的常规操作, 如噪声、压缩、格式转换等具有鲁棒性, 而对恶意攻击具有敏感性. 由于半脆弱数字水印技术能在一定程度上能区分保持内容的处理与恶意篡改操作^[1], 近些年来, 基于半脆弱数字水印技术的图像内容认证的研究得到了广泛的关注.

引用格式: 付剑晶, 陈德人, 徐达文, 等. 适用于双重压缩环境的图像内容认证水印算法. 中国科学: 信息科学, 2019, 49: 464–485, doi: 10.1360/N112018-00196
Fu J J, Chen D R, Xu D W, et al. A watermarking algorithm for image content authentication in double-compression environment (in Chinese). Sci Sin Inform, 2019, 49: 464–485, doi: 10.1360/N112018-00196

一般而言,用于内容篡改检测与定位的水印算法主要包括3个部分:基于内容特征的生成与重构、特征水印的嵌入与提取、恶意篡改的检测与定位.其中因内容特征是从载体数据中抽取出来的,如载体受攻击后保持语义,作为部分刻画载体语义的特征往往会呈现出较好的稳定性.而水印算法的鲁棒性由攻击后特征与水印信息一致性的概率来体现,因而,由水印嵌入算法体现的水印的鲁棒性就成了认证算法的关键,而且它还影响到有关检测的多项性能.

根据水印嵌入位置的不同,可将水印算法划分为两类^[2]:空域算法,即直接修改载体信号实现水印嵌入;变换域算法,也就是通过修改载体变换域系数来嵌入水印.一般而言,变换域的水印算法在鲁棒性和透明性上优于空域算法,所以基于数字水印的研究也主要集中在变换域.至今,有关数字水印的文献非常多(如文献[3~19]),为此领域的发展奠定了坚实的基础.

由于JPEG和JPEG2000压缩分别采用了DCT与DWT变换,很多基于DCT/DWT的水印算法相继出现,以减小压缩处理对水印信号的影响,如文献[3~15,18,19].在这些研究中,由于水印嵌入算法体现的水印都是在不同层面较低程度地反映载体数据的语义,所以这些算法在抵制攻击方面的性能,如压缩、噪声、滤波、亮度改变等,各有不同的侧重(限于论文篇幅,具体评述被略).

对内容认证水印算法而言,在满足不可感知性与容量的条件下,其鲁棒性更应得到重视;要全面提升水印算法鲁棒性的综合性能,设计能更大程度地体现载体语义的算法是个挑战.一些研究表明^[16,17],基于预测的算法有望进一步提升水印的鲁棒性.Wang等^[16]在空域和DCT域分别提出了具有一定鲁棒性信息隐藏算法,两种算法都是通过由邻域数据得到的中心数据预测值与实际值的大小关系来嵌入信息的.其中DCT域方法更鲁棒,能容忍较低品质的JPEG攻击,但对噪声、滤波较敏感,且容易引发明显的图像失真.Hsu和Hu^[17]利用基于相关调制和量化调制的混合调制技术,通过 5×5 网格中的12个数据块对中心块部分DCT系数的预测,建立了基于部分符号改变的均值量化水印算法(partly sign-altered mean modulation, PSAM).混合调制使该算法在不牺牲视觉质量条件下实现了很好的鲁棒性,而预测中体现的多个相对关系使其水印信号在一定程度上抓住了数据的语,从而能够抵制多种攻击.另一方面,为了提高算法的检测性能,文献[13]在检测阶段采用5个度量指标的3级处理方法,能在一定程度上弥补鲁棒性不够所导致的性能上的不足.

近些年来,一些研究将机器学习类算法,如神经网络、支持向量机等,引入到数字水印系统^[20,21],然而这些算法的计算开销导致的应用限制是个不容忽视的问题.基于机器学习的水印算法超出了本文的范围,有兴趣的读者可以去查阅.

在传统的数字水印研究领域,研究人员已做了大量工作,然而两类操作简单且极易出现的重要安全问题却经常被忽略.第1类是水印针对性的攻击,就是在保持载体使用价值不受影响的情况下,有针对性地对承载水印信号的数据进行破坏,会导致检测时的误判.这类问题源于水印的嵌入位置完全固定或相对固定,如文献[3,5,8~10,13,16~19]都存在此问题.第2类是拼贴攻击^[22,23],当水印序列与载体数据无关时,比如随机二进制序列和固定的Logo图像,对用相同的密钥和算法得到的两个水印载体 A_w 和 B_w ,如果将 B_w 部分数据块替换到 A_w 的对应数据块上便可形成伪造水印载体 F_w ,此时 F_w 可能逃避检测;此外,当数据块嵌入来自块本身的内容特征水印时,如将 A_w 中部分数据块替换到 A_w 中的其他数据块上可得到伪造水印载体 C_w ,此时 C_w 同样可能逃避检测,文献[3,5,7,8,10,11,13,15~18]都存这类问题.

然而至今,大多数水印研究,包括文献[3~19],还普遍存在如下问题.(1)假定水印载体在检测时至多经历一次压缩,而忽略了常规信号处理加一次压缩处理的常规应用组合情形^[3~19].(2)有关内容篡改的检测与定位主要是在无常规信号处理的条件下进行的,而未考虑篡改加常规信号处理的实际组合情形^[8,9,11,13].(3)难以解释含水印的数据是以何种方式去表达载体的语义的,往往表现为水印的表

示与载体语义无关或表达载体语义程度太低^[3,5~15,18],由此引发的问题是:(a) 水印的稳定性缺乏理论基础,鲁棒性事先难以评估;(b) 针对不同的攻击,算法在鲁棒性上波动较大.上述问题的存在,严重地影响了相关研究的应用推广.

此外,更值得重视的是,随着摄像传感器分辨率的提高,数字图像往往以有损压缩形式流通到应用领域,然后可能经历解码后的保持内容的信号处理和/或恶意篡改等操作,接着又以有损压缩形式(第2次压缩)存储或发布.如此在终端检测时,图像可能经历了两次压缩及期间的攻击(本文称之为双重压缩),且两次压缩常以 JPEG 和 JPEG2000 形成的 4 种组合情况存在.因此,研究能抗双重压缩的图像内容认证水印算法更具应用价值,这也对算法的鲁棒性提出了更高要求,至今相关水印文献尚未见报道.

本文基于旋转向量的方法,提出了一种新颖的水印调制算法,并在此基础上建立了半脆弱水印方案用于图像内容篡改检测与定位.由于水印调制算法使含水印的数据能在一定程度上体现载体数据的语义,从而使本文方案在双重压缩应用环境下能有效地区分保持内容的处理与恶意篡改,同时兼顾了水印的安全性.本文主要贡献如下:

- (1) 提出了根据含水印的数据以何种程度体现载体数据的语义来事先初步估计水印算法的鲁棒性的观点,有利于新算法探索中的可行性分析;
- (2) 提出了一种新颖的水印算法,包括基于旋转向量的水印表示方法与调制算法;
- (3) 理论分析了由水印数据的稳定性,相比于在已有的大多数水印文献中只给出算法及实验数据,本文的水印方案具有理论基础;
- (4) 提出的认证方案在双重压缩环境下能有效区分保持内容的处理与恶意篡改,并对双重压缩间的多种攻击,水印的鲁棒性好且分布稳定,推广了水印技术的应用范围;
- (5) 在双重压缩环境下的篡改定位性能在一定程度上能满足应用需求;
- (6) 在理论推导与实验数据基础上建立了相关性能度量模型,对方案的应用具有指导价值.

本文剩余部分内容介绍如下:第2节阐述了新颖的水印表示方法及调制算法,并对水印的稳定性进行了数学分析和实验测试;第3节描述了方案中的内容特征抽取算法,并在双重压缩条件下测试了特征的稳定性;第4节详细阐述了所提出的内容认证方案;第5节对方案中的相关性能进行了分析;第6节给出了实验结果与分析,并与相关方案进行了比较;第7节对全文进行了总结.

2 水印的表示与稳定性

在本节,我们先阐述本文方案中的水印表示方法及其对载体语义的体现,并从理论上分析其稳定性;然后给出详细的水印调制算法,最后通过实验来获取水印的具体稳定性能.

2.1 水印的表示

设 $B = \{b_{ij}\}_{m1 \times m2}$ 为载体中的数据块,在 B 的不同位置上随机选取 3 个数据构成向量 $\mathbf{A} = \{a_1, a_2, a_3\}$,如图 1 所示,在 3 维空间由向量 $\mathbf{A}$ 与给定的参考向量 $\mathbf{V}_r$ 所决定的平面 AOV_r 上,将向量 $\mathbf{A}$ 相对向量 $\mathbf{V}_r$ 正向靠近(图 1(a))或反向远离(图 1(b))旋转 θ 得到承载水印信息的向量 $\mathbf{A}_w$,使得由向量 $\mathbf{V}_r$ 与 $\mathbf{A}_w$ 形成的夹角 β 按式 (1) 所示的方式来体现数据块 B 中的水印信 w_B . 其中 θ_q 是

图 1 (网络版彩图) 水印向量的形成

Figure 1 (Color online) Watermarked vector $\mathbf{A}_w$ formed through rotating the original vector ($\mathbf{A}$) by θ . (a) Counterclockwise; (b) clockwise

已知的以角度为单位的量化步长, 且 $0 < \theta < \theta_q$.

$$w_B = \begin{cases} 1, & \lfloor \beta/\theta_q \rfloor \% 2 == 1, \\ 0, & \lfloor \beta/\theta_q \rfloor \% 2 == 0. \end{cases} \quad (1)$$

当向量在模长与空间指向上表达语义权重各占 50% 时, 由于 $\mathbf{A}_w$ 和 $\mathbf{A}$ 模长相等, $\mathbf{A}_w$ 能在 $0.5 + (180 - \theta)/180 \times 0.5$ 的程度上体现 $\mathbf{A}$ 所表达的语义; 由此 θ 取值越小, $\mathbf{A}_w$ 越能逼近 $\mathbf{A}$ 在 B 中所体现的语义. 然而在实际中, 向量 $\mathbf{A}$ 在 B 中所体现的语义, 不仅仅只包含模长和空间指向, 其语义内容与 B 的应用环境有关. 当 B 为图像数据块时, $\mathbf{A}$ 在 B 中所体现的语义主要表现在 $\mathbf{A}$ 中元素间的大小关系以及 $\mathbf{A}$ 中元素与 B 中其他元素的大小关系上; 此时, 因 $\mathbf{A}_w$ 由 $\mathbf{A}$ 旋转而形成, 当 θ 较小时, $\mathbf{A}$ 中元素间的大小关系在 $\mathbf{A}_w$ 中大部分得到保持. 因此, 从这个角度讲, $\mathbf{A}_w$ 能在一定程度上体现 $\mathbf{A}$ 在载体 B 中所表达的语义, 这也在后续的测试中得到了证实.

2.2 水印的稳定性分析

本小节在 2.1 小节的基础上分析水印的稳定性, 即分析向量 $\mathbf{A}_w$ 在保持内容处理的扰动下, 其空间指向的稳定性. 设 $\varepsilon = \{e_{ij}\}_{m1 \times m2}$ 为对水印载体 $B_w = \{c_{ij}\}_{m1 \times m2}$ 施加的一个扰动, 由于 ε 是保持内容的扰动, 故在一般情形下 ε 能量远小于 B_w 的能量, 即有 $\|\varepsilon\|_2 \ll \|B_w\|_2$. 另设 ε_w 为 ε 对向量 $\mathbf{A}_w$ 构成的扰动向量, 因此抖动后的水印向量为 $\mathbf{A}'_w = \mathbf{A}_w + \varepsilon_w$, 一般情形下同样有 $\|\varepsilon_w\|_2 \ll \|\mathbf{A}_w\|_2$.

设向量 $\mathbf{A}_w$ 与 $\mathbf{A}'_w$ 夹角为 θ_d , 即 $\mathbf{A}_w$ 抖动前后形成的抖动角, 并用 θ_ε 表示 ε_w 与 $\mathbf{A}_w$ 的夹角, 则

$$(\mathbf{A}_w, \varepsilon_w) = \|\mathbf{A}_w\|_2 \|\varepsilon_w\|_2 \cos(\theta_\varepsilon), \quad (2)$$

$$\cos(\theta_d) = \frac{(\mathbf{A}_w, \mathbf{A}_w + \varepsilon_w)}{\|\mathbf{A}_w\|_2 \times \|\mathbf{A}_w + \varepsilon_w\|_2} = \frac{\|\mathbf{A}_w\|_2^2 + (\mathbf{A}_w, \varepsilon_w)}{\|\mathbf{A}_w\|_2 \times \|\mathbf{A}_w + \varepsilon_w\|_2} \geq \frac{\|\mathbf{A}_w\|_2^2 + (\mathbf{A}_w, \varepsilon_w)}{\|\mathbf{A}_w\|_2 \times (\|\mathbf{A}_w\|_2 + \|\varepsilon_w\|_2)}, \quad (3)$$

于是, 根据式 (2) 和 (3),

$$\cos(\theta_d) \geq \frac{\|\mathbf{A}_w\|_2^2 + \|\mathbf{A}_w\|_2 \|\varepsilon_w\|_2 \cos(\theta_\varepsilon)}{\|\mathbf{A}_w\|_2 \times (\|\mathbf{A}_w\|_2 + \|\varepsilon_w\|_2)} = \frac{1 + \lambda \cos(\theta_\varepsilon)}{1 + \lambda} \geq \frac{1 - \lambda}{1 + \lambda}, \quad (4)$$

其中 $\lambda = \|\varepsilon_w\|_2 / \|\mathbf{A}_w\|_2$.

当 ε_w 与 $\mathbf{A}_w$ 同向或反向时, 有 $\varepsilon_w = \pm \lambda \mathbf{A}_w$, 因而

$$\cos(\theta_d) = \frac{(\mathbf{A}_w, \mathbf{A}_w + \varepsilon_w)}{\|\mathbf{A}_w\|_2 \times \|\mathbf{A}_w + \varepsilon_w\|_2} = \frac{(\mathbf{A}_w, \mathbf{A}_w(1 \pm \lambda))}{\|\mathbf{A}_w\|_2 \times \|\mathbf{A}_w(1 \pm \lambda)\|_2} = 1, \quad (5)$$

此时, $\theta_d = 0$.

根据式 (4) 和 (5), 归纳如下:

$$\begin{cases} \cos(\theta_d) = 1, & \varepsilon_w = \pm|\lambda|\mathbf{A}_w, \\ \frac{1-\lambda}{1+\lambda} \leq \cos(\theta_d) \leq 1, & \text{otherwise.} \end{cases} \quad (6)$$

式 (6) 表明, $\mathbf{A}_w$ 的空间指向在 ε_w 抖动下具有稳定性, 其稳定程度主要取决于 λ , 且与 $\mathbf{A}_w$ 的维度无关. 由于平凡情况下 $\|\varepsilon_w\|_2 \ll \|\mathbf{A}_w\|_2$, 即有 $\lambda \rightarrow 0$, 则 $\cos(\theta_d) \rightarrow 1$, 也就是 $\theta_d \rightarrow 0$.

同理, 对原始载体 $B = \{b_{ij}\}_{m1 \times m2}$ 中的向量 $\mathbf{A}$, 其在扰动 ε_A 下也具有稳定性. 由于 $\mathbf{A}_w$ 是由 $\mathbf{A}$ 旋转而来, 即 $\|\mathbf{A}\|_2 = \|\mathbf{A}_w\|_2$, 根据上述分析, 对相同的扰动, 向量 $\mathbf{A}$ 和 $\mathbf{A}_w$ 在空间指向上的稳定性程度是一致的.

2.3 水印调制方法

本小节在 2.1 小节的基础上, 对已知的 $B = \{b_{ij}\}_{m1 \times m2}$, $n = 3$, $\mathbf{A} = \{a_l\}_n$, θ_q , $\mathbf{V}_r = \{r_l\}_n$, 详细阐述计算 θ , β , $\mathbf{A}_w = \{x_l\}_n$ 的方法, 最终生成含水印信息的数据块 $B_w = \{c_{ij}\}_{m1 \times m2}$.

设向量 $\mathbf{A}$ 与 $\mathbf{V}_r$ 的夹角为 α , 则有

$$\alpha = \arccos\left(\frac{\mathbf{A} \cdot \mathbf{V}_r}{|\mathbf{A}||\mathbf{V}_r|}\right), \quad (7)$$

然后, 依据式 (1) 体现的反量化规则, 对待嵌入的水印信息 $w_B \in \{0, 1\}$, 通过量化的方式计算 β :

$$\beta = \begin{cases} \lfloor \alpha/\theta_q \rfloor \times \theta_q, & \lfloor \alpha/\theta_q \rfloor \% 2 == w_B, \\ \lfloor \alpha/\theta_q \rfloor \times \theta_q + \theta_q, & \lfloor \alpha/\theta_q \rfloor \% 2 \neq w_B. \end{cases} \quad (8)$$

由于 $\mathbf{A}_w$ 是由 $\mathbf{A}$ 在 AOV_r 平面上旋转而成的, 即向量 $\mathbf{A}$, $\mathbf{V}_r$, $\mathbf{A}_w$ 在同一平面上, 如图 1 所示, 因此有

$$\theta = \alpha - \beta. \quad (9)$$

为便于计算 $\mathbf{A}_w$, 本文采用 Rodrigue 旋转公式 (Rodrigues' rotation formula), 如下所示:

$$\beta = \begin{cases} \mathbf{A}_w = \cos(\theta) + (1 - \cos(\theta))(\mathbf{A} \cdot \text{norm}(\mathbf{V}_r)) \cdot \text{norm}(\mathbf{V}_r) + \sin(\theta)\text{norm}(\mathbf{V}_r) \times \mathbf{A}, \\ \text{norm}(\mathbf{V}_r) = \mathbf{V}_r/|\mathbf{V}_r|, \end{cases} \quad (10)$$

然后把向量 $\mathbf{A}_w$ 中的 x_1, x_2, x_3 分别替换向量 $\mathbf{A}$ 在载体 B 中对应的数据 a_1, a_2, a_3 , 便可得到含水印信息的载体 B_w .

2.4 水印稳定性测试

为观察水印在双重压缩条件下的稳定性的实际性能, 本小节将执行一系列的测试, 并且通过对测试数据的统计分析, 为本文方案在实际应用中量化步长 θ_q 的设置以及后续的性能分析提供参考数据. 由于 2.2 小节已表明水印向量 ($\mathbf{A}_w$) 与原向量 ($\mathbf{A}$) 在空间指向上的稳定性具有一致性, 为方便起见, 我们在原始载体数据上展开测试.

图 2 (网络版彩图) 向量 $\mathbf{A}$ 的抖动角度 (θ_d) 与累积概率 (CP) 的关系曲线

Figure 2 (Color online) The relationship between the dithered angle (θ_d) of vector $\mathbf{A}$ and the cumulative probability under perturbation for (a) test schemes 1~4, respectively, and (b) test scheme 5 and schemes 1~4

我们从 USC-SIPI image database¹⁾的 3 个类别 (Textures, Aerials, Miscellaneous) 中选取 50 张图片, 分辨率覆盖 256×256 , 512×512 , 1024×1024 ; 设计 13 种保持内容处理的单项攻击: 1 – 乘积性噪声 (0.002), 2 – Gauss 噪声 (0.002), 3 – 椒盐噪声 (0.002), 4 – 滤波均值 (3×3), 5 – 滤波中值 (3×3), 6 – 滤波维纳 (3×3), 7 – 均衡化 (164), 8 – 绝对变暗 (-30), 9 – 绝对增亮 ($+20$), 10 – 相对变暗 ($[0, 0.7]$), 11 – 相对增亮 ($[0.4, 1]$), 12 – 尺度放大并复原 (1.8×1.8), 13 – 尺度缩小并复原 (0.8×0.8). 测试在 Matlab 2016b 环境下执行, 其中有关亮度的相对改变 (图像对比度调整), 请参考 `imadjust(·)` 函数.

针对双重压缩, 设计如下 5 种测试方案: (1) JPEG($Q = 75$) $\rightarrow$ 单项攻击 $\rightarrow$ JPEG2000(Rate = 8); (2) JPEG2000(8) $\rightarrow$ 单项攻击 $\rightarrow$ JPEG(75); (3) JPEG(75) $\rightarrow$ 单项攻击 $\rightarrow$ JPEG(75); (4) JPEG2000(8) $\rightarrow$ 单项攻击 $\rightarrow$ JPEG2000(10); (5) 无双重压缩下的单项攻击. 其中, 方案 (1) 表明在两次压缩之间 (JPEG, JPEG2000) 只执行单项攻击 1~13 中的一项, 因每个单项攻击都要覆盖到, 所以共执行 13 次, 最后采用测试数据的平均值; 方案 (2)~(5) 与之类似.

根据 2.3 小节的内容, 设置 $n = 3$; 考虑到亮度的绝对调整, 即单项攻击 8 和 9 会破坏条件: $\|\varepsilon_w\|_2 \ll \|\mathbf{A}_w\|_2$ (2.2 小节), 因此对待检测数据块 B 先进行标准化预处理, 即设置 $B = B - \text{Avg}(B)$, 其中 $\text{Avg}(\cdot)$ 为求均值操作; 另外, 为减小双重压缩、噪声与滤波等对水印稳定性的影响, 我们把图像划分为 16×16 互不重叠的数据块, 并对每个数据块执行 2 级小波提升, 设置低频数据域 (LL2) 作为测试数据区, 即 $B = \{b_{ij}\}_{m1 \times m2}$ 为数据块标准化处理后的 LL2 域, 其中 $m1 = m2 = 4$. 这也表明本文方案中的水印嵌入与提取算法将在此区域进行. 最后, 向量 $\mathbf{A}$ 中的元素在密钥 k_1 控制下, 从 B 中不同位置随机选取.

向量 $\mathbf{A}$ 在攻击前后形成的抖动角度 (θ_d) 与累积概率 (CP) 的关系如图 2 所示. 图 2(a) 的曲线表明, 在双重压缩条件下, 向量的空间指向的稳定性总体表现良好, 其中方案 (4), 即双重 JPEG2000 压缩对向量稳定性的影响相对较大. 图 2(b) 为向量在未压缩和 4 种双重压缩总体情况下稳定性的对比图, 其

1) USC-SIPI image database. <http://sipi.usc.edu/database/>.

图 3 (网络版彩图) 双重压缩环境下的单项攻击的特征的稳定性

Figure 3 (Color online) The probability of feature stability for individual attack in double compression environment for (a) test schemes 1~4, respectively, and (b) test scheme 5 and schemes 1~4

中, $CP(\theta_d \leq 10 | \text{无压缩}) = 0.895$, $CP(\theta_d \leq 20 | \text{无压缩}) = 0.9798$; 而 $CP(\theta_d \leq 20 | \text{双重压缩}) = 0.8598$, $CP(\theta_d \leq 25 | \text{双重压缩}) = 0.8990$, $CP(\theta_d \leq 30 | \text{双重压缩}) = 0.9258$. 这些测试数据说明: (1) 本文基于旋转向量的水印表示方法是鲁棒的; (2) 提出的水印表示方法在双重压缩下具有可行性; (3) 为兼顾到鲁棒性与图像质量, 可考虑在区间 $[20, 40]$ 内设置量化步长 θ_q .

3 特征的抽取与稳定性

为抵制拼贴攻击 (详见第 1 节第 7 段), 我们把从图像数据块中抽取的内容依赖的特征作为水印, 并在密钥 k_2 的控制下将特征水印以交叉的方式嵌入到其他数据块中. 由于本文方案的水印嵌入与提取的数据域为数据块标准化后的二级小波域 (2.4 小节), 为了降低计算开销, 特征抽取也在此区域进行. 考虑到本文方案要抗击双重压缩, 而且 JPEG2000 压缩对小波变换的中高频系数影响较大, 因此特征抽取的区域进一步缩小到低频域 (LL2 域).

由于作为水印的承载数据, 向量 $\mathbf{A}$ 的元素是在密钥 k_1 的控制下, 从 B 的不同位置随机选取出来的; 为避免水印与特征的相互干扰, 我们在同样密钥的控制下, 从 B 的水印位置外的其他地方随机选择两个不同位置点 p_{f1} 和 p_{f2} , 然后根据这两个位置数据的大小关系来抽取数据块 B 的 1 比特内容特征, 如下所示:

$$f_B = \begin{cases} 1, & B(p_{f1}) \geq B(p_{f2}); \\ 0, & B(p_{f1}) < B(p_{f2}). \end{cases} \quad (11)$$

接下来, 通过一些测试来观察式 (11) 所示的内容特征的稳定性, 同时也为后续的性能分析提供参考数据. 测试环境与 5 种测试方案与 2.4 小节相同, 测试结果如图 3 所示. 图 3(a) 展示了 4 种双重压缩下单项攻击的特征稳定性, 稳定性概率在 $[0.91, 0.979]$ 区间内波动, 各方案都是在单项攻击 2 - Gauss 噪声 (0.002) 处, 即在以均值 0、标准差为 11.4 的 Gauss 噪声, 降到最低值. 与图 2(a) 情况相

同,在4种双重压缩中,压缩方案(4)对稳定性影响最大.在图3(b)中,未压缩情况下数据范围在区间 $[0.928, 1]$ 内;4种双重压缩总体统计数据落在区间 $[0.918, 0.961]$ 内,其均值为0.9492.这些测试数据说明式(11)所示的特征抽取方法在双重压缩环境下是完全可行的.

4 认证方案

在第2和3节基础上,本节对所提出的图像内容认证方案给出完整的描述.我们的认证方案首先把输入的图像按 16×16 的尺寸划分为互不重叠的子块,并对每个子块执行二级小波变换.然后在各子块的LL2域,根据密钥 k_1 随机选择位置互不相同的 $3+2$ 个数据,分别用于水印嵌入和特征抽取;其中从子块抽取的特征水印并不嵌入在当前块,而是受密钥 k_2 控制,以交叉的方式嵌入到其他数据块中.在检测端,通过密钥可以重构特征和提取水印,并形成关于各子块的特征与水印一致性矩阵;接着根据提出的篡改检测策略进行内容认证.本节主要内容包括:特征水印的嵌入、特征重构与水印提取、篡改检测.

4.1 特征水印的嵌入

为降低计算开销,本文方案的特征抽取与水印嵌入都在相同的变换域进行,而且为了进一步提高算法的效率,我们将二者融合到一个过程中进行处理,具体步骤如下.

(1) 子块划分和预处理.对给定的输入图像 I ,将其按尺寸 16×16 划分为互不重叠的子块,表示为 BK_i ($i = 1, 2, \dots, N$),其中 $N = \text{INT}(r_1/16) \times \text{INT}(r_2/16)$, $r_1 \times r_2$ 为 I 的分辨率;接着进行标准化处理以减小图像亮度绝对调整对水印性能的影响,即令 $\text{avg_BK}_i = \text{Avg}(BK_i)$, $BK_i = BK_i - \text{avg_BK}_i$,其中 $\text{Avg}(\cdot)$ 为求均值操作(2.4小节);然后对 BK_i 进行二级小波变换,得到二级低频域数据 $BK_{i,LL2}$ 和其他子带数据 $BK_{i,other}$.

(2) 密钥的生成.以 k_0 作为主密钥,随机产生3个数作为子密钥,分别表示为 k_1, k_2, k_3 ,以供后续过程使用.由此表明,本文所提出的水印方案只需一个密钥 k_0 即可.

(3) 水印与特征位置的确定.令 $B = BK_{i,LL2}$,根据子密钥 k_1 在 B 中随机选择 $n+2$ 个不相同的位置(本文设置 $n=3$),表示为 $p_{w1}, p_{w2}, \dots, p_{wn}$,和 p_{f1}, p_{f2} ;前 n 个位置的数据用来形成承载水印的向量 $\mathbf{A} = (a_1, a_2, \dots, a_n)$,且 $a_l = B(p_{wl})$, $l = 1, 2, \dots, n$; p_{f1} 和 p_{f2} 位置上对应的数据用于抽取特征.

(4) 特征抽取与交叉索引的建立.根据 B 中随机确定的两个已知位置 p_{f1} 和 p_{f2} ,和式(11)提取数据块的内容特征 f_B ,由此可形成整个图像的特征序列 $\mathbf{F} = (f_1, f_2, \dots, f_N)$.此外,根据子密钥 k_2 建立一维随机向量 $\mathbf{V} = (v_1, v_2, \dots, v_N)$,并对 $\mathbf{V}$ 进行升序排序,得到索引向量 $\mathbf{C}_{\text{idx}} = (c_1, c_2, \dots, c_N)$,其中 c_i 为1到 N 间的整数.

(5) 参考向量的生成.以 k_3 为子密钥,随机生成维度为 n 的随机参考向量 $\mathbf{V}_r$.

(6) 特征水印的交叉嵌入.令 $B = BK_{i,LL2}$,根据步骤(3)和(5)可得到已知量 $\mathbf{A}$, n , $\mathbf{V}_r$,并令 $w_B = \mathbf{F}(\mathbf{C}_{\text{idx}}(i))$;此外,量化步长 θ_q 可参照图2(b)所示得到的测试数据和实际应用情况来确定,在本文,我们设置 $\theta_q = 30$ ($0 < \theta_q < 90$).接着按2.3小节的内容完成水印嵌入,得到含水印的数据 B_w ;然后根据 B_w , $BK_{i,other}$ 执行二级小波逆变换,得到含水印的数据 $BK_{i,w}$;最后执行 $BK_{i,w} = BK_{i,w} + \text{avg_BK}_i$,由此生成数据块 BK_i 对应的水印数据块 $BK_{i,w}$.

(7) 水印载体的输出.对每个数据块 BK_i 反复执行步骤(6),最后输出含水印的图像 I_w .

4.2 特征重构与水印提取

特征重构与水印提取过程是 4.1 小节反向过程. 对待检测图像 I' , 执行 4.1 小节中的步骤 (1)~(4), 可得到重构的特征向量 $\mathbf{F}'$, 表示为

$$\mathbf{F}' = (f'_1, f'_2, \dots, f'_N). \quad (12)$$

根据 2.3 小节式 (8), 单个数据块 B' 的 1 比特水印 $w_{B'}$ 按下式提取:

$$w_{B'} = \text{round}(\alpha' / \theta_q) \% 2, \quad (13)$$

其中 α' 是基于式 (7) 从 I' 中得到的. 对 I' 的每个数据块 B' 按照式 (13) 提取水印, 可得到水印向量 $\mathbf{W}'_k = ((w'_1)^k, (w'_2)^k, \dots, (w'_N)^k)$, 然后对 $\mathbf{W}'_k$ 执行以子密钥为 k_2 的反置乱 (4.1 小节步骤 (4)), 得到各数据块反交叉后的水印向量 $\mathbf{W}'$:

$$\mathbf{W}' = (w'_1, w'_2, w'_3, \dots, w'_N). \quad (14)$$

4.3 篡改检测

设待检测的图像 I' 的分辨率为 $r_1 \times r_2$, 我们先把重构的特征 $\mathbf{F}'$ 和提取的水印 $\mathbf{W}'$ 转换为矩阵形式, 以便于能与图像数据块直接对应起来. 转换如下:

$$\begin{cases} W^*(i, j) = \mathbf{W}'(k), & i = (k-1)/(r_2/16) + 1, \\ F^*(i, j) = \mathbf{F}'(k), & j = (k-1)\%(r_2/16) + 1. \end{cases}$$

其中, i, j, k 为整数, 且 $1 \leq k \leq N$, $1 \leq i \leq r_1/16$, $1 \leq j \leq r_2/16$. 检测步骤如下:

(1) 建立特征与水印一致性矩阵. 针对每个数据块, 通过逐个比较重构的特征与提取的水印, 建立一致性矩阵 T^0 :

$$\begin{aligned} T^0 &= \{t^0_{i,j} \mid 1 \leq i \leq r_1/16, 1 \leq j \leq r_2/16\}, \\ t^0_{i,j} &= \begin{cases} 0, & W^*(i, j) = F^*(i, j), \\ 1, & W^*(i, j) \neq F^*(i, j). \end{cases} \end{aligned} \quad (15)$$

(2) 生成篡改矩阵. 由于双重压缩保持内容的处理会使 T^0 中的“1”呈现随机离散均匀分布的现象, 而篡改往往是在局部区域集中进行的, 这又使 T^0 中的“1”可能呈现局部簇状. 为了过滤 T^0 中因保持内容处理操作形成的随机噪声, 对 T^0 中的每个点为中心的 3×3 的 9 邻域进行统计, 若标明为“1”的点至少有 5 个, 则该点标记为“1”, 表明为篡改块, 否则认为该点由保持内容的处理临时产生的, 标记为“0”, 表明非篡改块. 根据式 (15), 篡改矩阵 T 按下式生成:

$$\begin{aligned} T &= \{t_{i,j} \mid 1 \leq i \leq r_1/16, 1 \leq j \leq r_2/16\}, \\ t_{i,j} &= \begin{cases} 1, & \delta_{i,j} \geq 5, \\ 0, & \text{otherwise,} \end{cases} \quad \delta_{i,j} = \sum_{k=i, i \pm 1, l=j, j \pm 1} t^0_{k,l}. \end{aligned} \quad (16)$$

(3) 篡改检测. 为了客观判定待检测的图像 I' 是否含水印, 以及其内容是否遭恶意篡改, 定义比特错误率 (BER, Bit error rate) 和篡改簇比率 (TCR, tampering cluster rate), 分别如下所示:

$$\text{BER} = \frac{256 \times \sum_{i=1}^{r_1/16} \sum_{j=1}^{r_2/16} t^0_{i,j}}{r_1 \times r_2}, \quad (17)$$

$$\text{TCR} = \left(\sum_{i=1}^{r_1/16} \sum_{j=1}^{r_2/16} t_{i,j} \right) / \left(\sum_{i=1}^{r_1/16} \sum_{j=1}^{r_2/16} t_{i,j}^0 \right). \quad (18)$$

检测过程如下:

- (1) 当 $\text{BER} > \tau_1$, 图像不含水印, 不能检测;
- (2) 当 $\text{BER} \leq \tau_1$, 若 $\text{TCR} \leq \tau_2$ 图像内容可信;
- (3) 当 $\text{BER} \leq \tau_1$, 若 $\tau_2 < \text{TCR} \leq \tau_3$, 图像内容基本可信, 可能经历强度很大的信号处理, 或遭遇比例很小的内容篡改;
- (4) 当 $\text{BER} \leq \tau_1$, 若 $\text{TCR} > \tau_3$, 图像内容不可信, 图像内容遭恶意篡改.

其中, τ_1, τ_2, τ_3 为事先确定的与应用环境相关的检测阈值, 而且 τ_3 的取值又直接与 τ_2 有关. 由于这 3 个阈值的事先设定非常重要, 在后续内容中, 将在 2.4 小节与第 3 节的测试数据基础上, 结合相关推导, 给出其应用参考值 (详见 5.2 和 5.3 小节). 检测时若出现情形 (3) 和 (4) 时, 则根据篡改矩阵 T 中的“1”在检测图像的对应位置标记出来, 以供后续应用参考.

5 性能分析

本节根据在双重压缩环境下获得的测试数据 (2.4 小节、第 3 节), 针对所提出的认证方案进行有关性能分析, 内容包括鲁棒性分析 (水印与特征一致性概率)、相关检测阈值的设定及性能分析、内容篡改发生时的定位性能分析和安全性分析 (水印针对性攻击、拼贴攻击、伪造攻击).

5.1 鲁棒性

图像经历双重压缩及期间的信号处理后, 我们把从数据块 B' 中提取的水印 $w_{B'}$ 与重构的特征 $f_{B'}$ 的一致性概率称为鲁棒性, 表示为 $P(w_{B'} = f_{B'})$, 如下所示:

$$\begin{aligned} P(w_{B'} = f_{B'}) &= P(w_{B'} = w_B) \times P(f_{B'} = f_B) + P(w_{B'} \neq w_B) \times P(f_{B'} \neq f_B) \\ &= P(w_{B'} = w_B) \times P(f_{B'} = f_B) + (1 - P(w_{B'} = w_B)) \times (1 - P(f_{B'} = f_B)). \end{aligned} \quad (19)$$

由于本文双重压缩保持内容的处理过程复杂, 参数众多, 难以建立 $P(w_{B'} = w_B)$ 与 $P(f_{B'} = f_B)$ 的解析表达, 在此我们根据测试数据计算二者在平凡情况下的取值. 由式 (8) 与 (13) 决定的量化与反量化规则可得

$$P(w_{B'} = w_B) = \sum_k P\{2k\theta_q - \theta_q/2 \leq \theta_d \leq 2k\theta_q + \theta_q/2\}, \quad k = 0, 1, 2, 3, \dots, \quad 2k\theta_q + \theta_q/2 \leq 180. \quad (20)$$

根据 2.4 小节图 2(b) 所示的测试数据, 当 $\theta_q = 30$ 时有

$$\begin{aligned} P(w_{B'} = w_B) &= \text{CP}\{0 \leq \theta_d \leq 15\} + \text{CP}\{45 \leq \theta_d \leq 75\} + \text{CP}\{105 \leq \theta_d \leq 135\} \\ &\quad + \text{CP}\{165 \leq \theta_d \leq 180\} = 0.8253. \end{aligned} \quad (21)$$

再通过第 3 节得到的测试数据有

$$P(f_{B'} = f_B) = 0.9492, \quad (22)$$

然后把式 (21) 和 (22) 代入式 (19), 便得到鲁棒性的期望值, 如下式所示:

$$E[P(w_{B'} = f_{B'} | \theta_d = 30, \text{double compression})] = 0.7922. \quad (23)$$

近年来, 有关数字图像水印的文献非常多, 且相关算法对压缩的鲁棒性都只是考虑了单重压缩 (JPEG 或 JPEG2000) 情形; 而本文提到的双重压缩保持内容的处理是水印载体在应用中经常会遇到的情况, 因它对载体的信号损伤较大, 水印能经受单重压缩处理而存活的往往经受不住双重压缩. 式 (23) 是后续性能评价的基础, 也充分说明本文方案适用于双重压缩环境.

5.2 水印检测性能

作为检测过程的第 1 步, 就是根据阈值 τ_1 判断载体数据是否含有水印, 其实际设定值要视应用环境适当调整, 接下来给出其理论参考值.

在式 (15) 所示的矩阵 T^0 中, 每个点出现 “0” 或 “1” 的事件可以看作一组随机变量 ξ_k ($k = 1, 2, \dots$), 它们相互独立且有相同的 0-1 分布; 其中 $P(\xi_k = 1) = p$, $P(\xi_k = 0) = q = 1 - p$. 令 $\eta_n = \sum_{k=1}^n \xi_k$, 则 $\eta_n \sim B(n, p)$, 且有 $E(\eta_n) = np$, $D(\eta_n) = npq$. 根据 DeMoivre-Laplace 中心极限定理, 当 p 不变, n 无限增大时, η_n 以标准正态分布为其极限分布, 即 $\eta_n \sim N(0, 1)$.

我们用 $P_{\tau_1, \text{FP}}$ 和 $P_{\tau_1, \text{FN}}$ 分别表示 τ_1 的虚警率 (无水印载体被判为有水印的概率) 和漏警率 (载体含水印被判为无水印的概率), 则有

$$P_{\tau_1, \text{FP}} = P(\text{BER} > \tau_1 | \text{watermark}) = 1 - P(\text{BER} \leq \tau_1 | \text{watermark}) \approx \Phi\left(\frac{x - np}{\sqrt{npq}}\right), \quad (24)$$

$$P_{\tau_1, \text{NP}} = P(\text{BER} \leq \tau_1 | \text{no watermark}) \approx \Phi\left(\frac{n\tau_1 - 0.5n}{\sqrt{n \times 0.25}}\right), \quad (25)$$

其中, $\Phi(\cdot)$ 为正态分布函数, $n = \text{INT}(r_1/16) \times \text{INT}(r_2/16)$, $r_1 \times r_2$ 为 I 的分辨率, 且 $x = n \times \tau_1$; 此外, 根据式 (23) 有 $p = 1 - 0.7922 = 0.2018$.

当 $r_1 = r_2 = 512$ 时, 有 $n = 1024$; 我们对 τ_1 设置 1000 个等间隔点, 即 $\tau_1 = 0 : 0.001 : 1$, 然后根据式 (24) 和 (25) 计算, 便可得到 τ_1 分别与 $P_{\tau_1, \text{FP}}$ 和 $P_{\tau_1, \text{FN}}$ 的关系曲线图 (限于篇幅, 图被略). 本文取 $\tau_1 = 0.385$ 作为应用的参考, 此时有虚警率 $P_{\tau_1, \text{FP}} = 6.8401 \times 10^{-14}$ 和漏警率 $P_{\tau_1, \text{FN}} = 0.0$.

5.3 篡改检测性能

式 (18) 中的篡改簇比率 TCR 是来自篡改矩阵 T 中 “1” 的数量与初始矩阵 T^0 中 “1” 的数量的比率, 当它超过一定值时可反映图像被篡改的情况. 我们用 $P_{\tau_2, \text{FP}}$ 和 $P_{\tau_2, \text{FN}}$ 分别表示 τ_2 的虚警率 (含水印的载体未遭篡改而判为篡改) 和漏警率 (含水印的载体遭篡改而判为无篡改).

矩阵 T 中每个点出现 “0” 或 “1” 的事件同样可以看作 ξ_k ($k = 1, 2, \dots$) 相互独立且有相同的 0-1 分布, 与 5.2 小节类似, 设 $P(\xi_k = 1) = p_1$, $P(\xi_k = 0) = q_1 = 1 - p_1$, 并设 p_0 代表篡改发生时矩阵 T^0 中出现 “1” 的概率, 则 p_0 与篡改比例有关; 为了简化问题, 基于式 (23), 设定平均篡改率 $p_0 = [(1 - 0.7922) + 0.5]/2 = 0.3539$, 于是有

$$P_{\tau_2, \text{FN}} = P(\text{TCR} < \tau_2 | \text{be tampered}) \approx \Phi\left(\frac{x_1 - np_1}{\sqrt{np_1q_1}}\right), \quad (26)$$

其中 $x_1 = n \times p_0 \times \tau_2$, 并根据式 (16) 能得到在篡改发生时 T 中水印与特征不一致的概率 p_1 , 如下式所示:

$$p_1 = \sum_{k=5}^9 C_9^k \times p_0^k \times (1 - p_0)^k = 0.1784. \quad (27)$$

同时有

$$\begin{aligned} P_{\tau_2_FP} &= P(\text{TCR} > \tau_2 \mid \text{not be tempered}) = 1 - P(\text{TCR} \leq \tau_2 \mid \text{not be tempered}) \\ &\approx \Phi\left(\frac{x_2 - np_2}{\sqrt{np_2(1 - p_2)}}\right), \end{aligned} \quad (28)$$

其中, $x_2 = n \times p \times \tau_2$, 并且基于式 (16) 计算未篡改情况下矩阵 T 中出现“1”概率 p_2 如下:

$$p_2 = \sum_{k=5}^9 C_9^k \times p^k \times (1 - p)^k = 0.023. \quad (29)$$

当 $n = 1024$ 时, 我们也对 τ_2 设置 1000 个等间隔点, 即 $\tau_2 = 0 : 0.001 : 1$, 然后根据式 (26)~(29) 计算, 能得到 τ_2 分别与 $P_{\tau_2_FP}$ 和 $P_{\tau_2_FN}$ 的关系曲线图 (限于篇幅, 图被略). 为折衷 $P_{\tau_2_FP}$ 和 $P_{\tau_2_FN}$ 的性能, 应使二者的距离尽量小; 本文设置 $\tau_2 = 0.263$, 此时 $P_{\tau_2_FP} = 8.7182 \times 10^{-12}$, $P_{\tau_2_FN} = 4.2947 \times 10^{-13}$.

由于应用环境涉及到载体分辨率、不同信号处理的多种参数、双重压缩强度、篡改比例等众多事先不可确定的因素, 为降低误判率, 应在 τ_2 的基础上通过适当地上调来设置 τ_3 .

5.4 篡改定位性能

水印载体在经历保持内容的处理 (双重压缩) 后, 若内容又遭受恶意篡改, 我们将根据用矩阵 T 中“1”的位置来定位篡改数据块. 本小节基于篡改比例 p_t ($0 \leq p_t \leq 1$) 来分析矩阵 T 中“1”出现时的虚警率 (真实块判为篡改块的概率) 和漏警率 (篡改块判为真实块的概率), 分别表示为 P_{tl_fp} 和 P_{tl_fn} ; 并分析篡改矩阵 T 的鲁棒性 (T 中“0”出现的概率), 记为 P_{t_tr} ; 最后分析篡改定位准确率的理论值 (矩阵 T 中“1”出现时的准确率), 记为 P_{tl_d} .

首先把检测数据块划分为真实数据区 (记为 D_0) 和篡改数据区 (记为 D_1), 则对任意被检测数据块 B' 有

$$\begin{aligned} P\{B' \in D_1\} &= p_t, \\ P\{B' \in D_0\} &= 1 - p_t. \end{aligned} \quad (30)$$

为简化分析过程, 假定待检数据块的 9 邻域都属于同一数据区, 即忽略跨界情况, 由于本文的特征水印是采用交叉的方式来嵌入的, 因此在计算重构的特征与水印的不一致性概率时有如下 3 种情况:

$$P(w_{B'} \neq f_{B'}) \begin{cases} 0.5, & (B' \in D_0 \wedge f_{B'} \in D_1), \\ 1 - P(w_{B'} = f_{B'}), & (B' \in D_0 \wedge f_{B'} \in D_0), \\ 0.5, & B' \in D_1. \end{cases} \quad (31)$$

因此有

$$\begin{aligned} P(w_{B'} \neq f_{B'} \mid B' \in D_0) &= P(w_{B'} \neq f_{B'} \mid B' \in D_0 \wedge f_{B'} \in D_0) \times (1 - p_t) \\ &\quad + P(w_{B'} \neq f_{B'} \mid B' \in D_0 \wedge f_{B'} \in D_1) \times p_t \end{aligned}$$

$$= [1 - P(w_{B'} = f_{B'})] \times (1 - p_t) + 0.5p_t, \quad (32)$$

$$P(w_{B'} \neq f_{B'} | B' \in D_1) = 1 - P(w_{B'} = f_{B'} | B' \in D_1) = P(w_{B'} = f_{B'} | B' \in D_1) = 0.5. \quad (33)$$

再联合式 (31)~(33) 有

$$\begin{aligned} P(w_{B'} \neq f_{B'}) &= P(w_{B'} \neq f_{B'} | B' \in D_0) \times (1 - p_t) + P(w_{B'} \neq f_{B'} | B' \in D_1) \times p_t \\ &= \{[1 - P(w_{B'} = f_{B'})](1 - p_t) + 0.5p_t\}(1 - p_t) + 0.5p_t. \end{aligned} \quad (34)$$

然后, 根据式 (16), (31)~(34) 可分别建立对 P_{tl_fp} , P_{tl_fn} 和 P_{t_Tr} 的推导, 如式 (35)~(37) 所示:

$$\begin{aligned} P_{tl_fp} &= P(t_{i,j(B')} = 1 | B' \in D_0) = P(\delta_{i,j(B')} \geq 5 | B' \in D_0) \\ &= \left\{ 1 - \sum_{k=0}^{5-1} C_9^k P(w_{B'} \neq f_{B'} | B' \in D_0)^k (1 - P(w_{B'} \neq f_{B'} | B' \in D_0))^{9-k} \right\} \times (1 - p_t), \end{aligned} \quad (35)$$

$$\begin{aligned} P_{tl_fn} &= P(t_{i,j(B')} = 0 | B' \in D_1) = P(\delta_{i,j(B')} \geq 5 - 1 | B' \in D_1) \\ &= \left\{ \sum_{k=0}^{5-1} C_9^k P(w_{B'} \neq f_{B'} | B' \in D_1)^k (1 - P(w_{B'} \neq f_{B'} | B' \in D_1))^{9-k} \right\} \times p_t \\ &= \left\{ \sum_{k=0}^{5-1} C_9^k P(w_{B'} \neq f_{B'} | B' \in D_1)^9 \right\} \times p_t, \end{aligned} \quad (36)$$

$$P_{t_Tr} = P(t_{i,j(B')} = 0) = P(\delta_{i,j(B')} < 5) = \sum_{k=0}^{5-1} C_9^k P(w_{B'} \neq f_{B'})^k (1 - P(w_{B'} \neq f_{B'}))^{9-k}, \quad (37)$$

此外, 在式 (35) 和 (36) 的基础上可得到篡改定位的准确率 P_{tl_d} :

$$P_{tl_d} = 1 - [P_{tl_fp} \times (1 - p_t) + P_{tl_fn} \times p_t]. \quad (38)$$

最后, 通过把式 (23) 代入到式 (32)~(34), 再把式 (32)~(34) 代入到式 (35)~(38), 即可得到篡改比例 p_t 分别与鲁棒性 P_{t_Tr} 、篡改定位虚警率 P_{tl_fp} 、漏警率 P_{tl_fn} 、准确率 P_{tl_d} 的函数关系, 如图 4 所示. 当 p_t 为 0.0 与 0.2 时, 鲁棒性、定位虚警率、漏警率、准确率分别为 0.9770, 0.0230, 0.0, 0.9770, 和 0.8895, 0.0489, 0.0960, 0.9420. 虽然篡改比例 p_t 在实际应用中是事先未知的, 然而本小节内容的意义在于既能从理论上展示篡改矩阵 T 中出现“0”或“1”的相关性能, 也能为方案应用前的评估及应用后的进一步诊断提供辅助信息.

为实现内容认证中的篡改定位功能, 目前普遍采用的方法是在单个数据块中嵌入 1 比特鲁棒性水印, 然而这会导致在篡改区域中, 定位的漏警率约为 0.5, 也就是约有半数的篡改块不能被发现. 因此要降低定位的漏警率, 往往会考虑在单个数据块中增加鲁棒性水印的容量, 但这又使得经多比特特征与水印一致性体现的单数据块的鲁棒性会下降, 从而导致虚警率的升高, 除非 1 比特特征与水印的一致性概率在双重压缩环境下接近 1.0. 如何在水印的透明性、容量、鲁棒性三者的相互制约中解决这个问题, 有待进一步研究.

图 4 (网络版彩图) 双重压缩环境下, 篡改比例 p_t 分别与鲁棒性 ($P_{t.Tr}$), 定位的虚警率 ($P_{t.l.fp}$)、漏警率 ($P_{t.l.fn}$)、准确率 ($P_{t.l.d}$) 的函数关系

Figure 4 (Color online) Function relationship between tampering proportion (p_t) and robustness of the tampering matrix ($P_{t.Tr}$), false positive probability for localization ($P_{t.l.fp}$), false negative probability for localization ($P_{t.l.fn}$) and tampering localization accuracy ($P_{t.l.d}$), respectively, under double

5.5 安全性

(1) 水印针对性攻击. 由于本文方案的特征与水印位置是在数据块的二级小波低频域中, 从 16 个位置随机选取 5 个位置形成的, 在未知密钥 k_0 时, 不能对特征、水印, 以及二者同时实施针对性的攻击.

(2) 拼贴攻击. 在本文方案生成的水印载体中, 因当前数据块嵌入的水印来自密钥 k_2 决定的其他数据块的内容依赖的特征, 所以就单个拼贴块而言, 其水印与特征一致性概率为 0.5, 当拼贴块局部集中且达到一定数量时 (与 τ_2 有关), 则逃避不了终端的检测.

(3) 伪造攻击. 这类攻击旨在避开检测的同时又能按意图实施内容篡改, 但是有两个前提条件: (a) 已知被替换数据块的特征与水印的位置; (b) 根据公开的水印算法能使篡改块的特征和水印都与被替换块一致. 根据 4.1 小节的内容, 在未知密钥 k_0 时, 就不能明确特征与水印的位置, 并且也不能获得参考向量, 因此若要对本文方案生成的水印载体实施伪造攻击, 这两个前提条件是不成立的.

6 实验结果

为验证本文水印方案的透明性、鲁棒性和双重压缩环境下的篡改检测性能, 我们执行了一系列的实验, 设置如下: (1) 图像材料, 包括 8 幅具有代表性的 8-bit 512×512 灰度图像 (Lena, Sailboat, Baboon, Peppers, Man, F16, Boat, Cameraman), 以及 120 幅从 USC-SIPI image database 和 ILSVRC2016 dataset²⁾中下载的图像; (2) 图像攻击, 采用 2.4 小节中的 13 种单项攻击; (3) 双重压缩方案, 包括 JPEG 与 JPEG2000 的 4 种不同的组合与无压缩的情形, 详见 2.4 小节的 5 种测试方案; (4) 水印嵌入与提取参数, 设置量化步长 $\theta_q = 30$; (5) 测试环境, 采用 MATLAB 2016b 64 位编程环境.

相比于基于机器学习类的水印算法^[20, 21], 本文算法属于传统的水印算法, 而且有关抗双重压缩的图像水印文献至今未见报道; 在此情形下, 为尽可能公平地将本文方案与其他方案进行比较, 我们选取了 3 种相关的传统水印算法: (1) 由 Wang 等^[16]提出的基于 DCT 域系数相关性 (self-reference, SR)

2) ILSVRC dataset. <http://image-net.org/challenges/LSVRC/2016/download-images-8r28.php>.

表 1 峰值信噪比 (PSNR) 比较 ^{a)}
Table 1 Comparison of PSNR values

	Lena	Sailboat	Baboon	Peppers	Man	Average (128 images)
Ours	40.71	39.03	42.46	39.80	39.41	41.07
Qi	40.28	40.30	40.35	40.33	40.32	40.30
Wang	34.13	31.40	31.58	34.84	32.53	34.00
Hsu	38.69	37.38	37.78	38.45	37.80	38.19

a) The bold font represents the maximum of the corresponding column.

表 2 结构相似度 (SSIM) 比较 ^{a)}
Table 2 Comparison of SSIM values

	Lena	Sailboat	Baboon	Peppers	Man	Average (128 images)
Ours	0.9861	0.9885	0.9960	0.9837	0.9862	0.9897
Qi	0.9598	0.9707	0.9866	0.9609	0.9704	0.9701
Wang	0.9620	0.9527	0.9659	0.9653	0.9480	0.9627
Hsu	0.9563	0.9629	0.9826	0.9554	0.9631	0.9650

a) The bold font represents the maximum of the corresponding column.

算法, (2) 由 Qi 等^[13]提出的基于 DWT 域单系数自适应量化的算法, (3) 由 Hsu 等^[17]提出的结合量化调制与相关调制优势的 DCT 域 PSAM-MM 算法. 有关图像的传统水印文献非常多, 之所以选择上述 3 种算法进行比较, 出于以下考虑: (1) 本文算法与 Wang 和 Hsu 的算法都利用了变换域系数的相关性, 使水印数据能在一定程度上体现载体数据的语义, 从而获得较好的鲁棒性; (2) 据我们所知, Hsu 提出的算法是此前传统类水印算法中, 综合鲁棒性最好的水印算法; (3) 相比于 Wang 和 Hsu 的算法, 本文与 Qi 的算法都提出了用于篡改检测与定位的策略, 目标都是用于图像内容级认证, 此外, Qi 提出的 5 种度量制与 3 级检测策略, 对于内容认证领域的研究具有很好的参考价值.

在实验中, 相关算法的参数设置如下: 对 Wang 氏算法, 选取 $AC(1, 1)$ 嵌入水印, 并令 $\Delta = 0.05 \times |AC(1, 1)|$; 对 Hsu 氏算法, 采用与该文实验部分相同的参数; 至于 Qi 氏算法, 块尺寸采用文中提及的 4×4 和 8×8 的前者, 量化参数 q 设定为文中指定的自适应集 $\{11, 13, 15, 17\}$ 中的最大值, 以便突出该算法的鲁棒性.

6.1 水印的透明性

本小节选取峰值信噪比 (PSNR) 和结构相似度 (SSIM)^[24] 作为度量指标, 将本文算法及 3 种相关算法 (Qi, Wang, Hsu) 产生的水印图像进行客观评价 (在主观评价中, 除 Wang 氏算法会产生明显块痕迹外, 其余三者无明显差异, 限于篇幅, 相关内容在此被略).

表 1 和 2 就 PSNR 和 SSIM 分别给出了对 5 幅图像和 128 幅图像测试的详细数据和均值数据. 从表 1 数据可知, 本文方案的水印透明性略优于 Qi 氏和 Hsu 氏算法, 原因归纳如下: (1) 本文算法为抵制双重压缩, 需突出鲁棒性, 所以适当地牺牲了水印容量而采用了较大的块尺寸 (16×16); (2) 在为嵌入水印而修改的像素比例中, 本文算法最小, 只有 3/16.

6.2 水印的鲁棒性

为比较本文算法与其他 3 种相关算法的鲁棒性, 本小节针对单独的 JPEG 压缩、JPEG2000 压缩, 以及 5 种双重压缩方案 (包括无压缩下的图像攻击, 详见 2.4 小节), 执行了一系列攻击实验. 度量指

图 5 (网络版彩图) 单重压缩的鲁棒性

Figure 5 (Color online) Robustness of single compression. (a) JPEG; (b) JPEG2000

标采用比特正确率 (BCR, Bit correction rate), 即水印嵌入与提取的一致性数量与水印总数的比率; 所有算法均采用 0-1 随机序列作为水印。

本文对 128 幅图像分别执行了 9 级的压缩攻击实验, 结果如图 5 所示, 其中图例中的数字是 BCR 均值。由图 5 可知, 本文与 Hsu 氏算法在 BCR 均值上优于其他 2 种算法, 且 Hsu 氏算法性能最好; 在 JPEG 质量因子为 10% 处, 本文算法性能优于 Hsu 氏算法。

由于本文的目标是要在双重压缩环境下实现内容级认证, 因此双重压缩环境下水印算法的鲁棒性就成为了整个认证方案的重点。图 6 详细地展示本文算法和 3 个相关算法在双重压缩环境期间, 经不同的图像攻击后的鲁棒性数据。从图 6 可知, 在双重压缩环境下, 本文与 Hsu 氏算法在 BCR 均值上优于其他 2 种算法; 在无压缩情形, 本文算法性能优于 Hsu 氏算法 (图 6(e)), 而在 4 种压缩的总体情况下 (图 6(f)), 本文算法性能略低于 Hsu 氏算法。此外, 从图 6(a)~(f) 中的单项图像攻击性能分布上看, 相比于 Hsu 氏算法, 本文算法经不同攻击后的 BCR 值波动幅度较小, 在分布上呈现出更好的稳定性。这说明本文方案的水印数据在较大程度保持了原有数据多对多的关系, 相比于 Hsu 氏采用的多对一的关系, 我们的水印数据能更好地体现载体数据的语义。综上比较与分析, 就抗攻击的普遍适用性而言, 本文算法比 Hsu 氏算法更实用。

6.3 篡改检测与定位

本小节的目标是测试提出的认证方案在双重压缩环境下对保持内容的处理与恶意篡改的区分能力, 及其篡改定位性能。由于在 Wang 氏和 Hsu 氏的方案中都未给出篡改检测策略, 因而二者更适用于版权保护; 而对 Qi 氏方案, 虽然阐述了详细篡改检测策略, 但因其鲁棒性不适用于双重压缩环境 (6.2 小节), 因此本小节就不提供相关的比较结果。

根据 4.3, 5.2 和 5.3 小节的内容, 我们设定水印检测阈值 $\tau_1 = 0.385$, 篡改检测阈值 $\tau_2 = 0.263$ 。而 τ_3 应根据 τ_2 和应用环境适当取值; 若 τ_3 取值较大, 则将不可信图像误判为“基本可信”的概率会增加; 反之, 将基本可信图像误判为“内容不可信”的概率也会增加; 在本小节实验中, 我们设定 $\tau_3 = 0.333$ 。

图 7 和 8 分别给出了双重压缩环境下保持内容的图像攻击与恶意内容篡改后的检测蒙板图; 其中的子图题给出了图像经历的信号处理过程。例如, 图 7(j) 的子图题“DC3 $\leftarrow$ A8”表示该图像在执行双重压缩方案 (3) 期间, 采用攻击编号为 8 (8 - 绝对变暗 (-30)) 的图像处理; 图 8(k) 的子图题“DC4 $\leftarrow$ TA11”表示该图像在执行双重压缩方案 (4) 期间, 依次执行了内容篡改 (“T”) 和图像攻击 11 (11 - 相对增亮 ([0.4, 1])). 此外, 图 8 中的白色矩形框标示出了相应的篡改区域, 其中子图 (a)~(d) 采用了

图6 (网络版彩图) 在双重压缩间不同图像攻击的鲁棒性

Figure 6 (Color online) Robustness of different image attacks during double compression. (a)~(d) correspond to the double compression schemes 1~4, respectively, (e) is the uncompressed case, and (f) is the average of (a)~(d)

拼贴篡改 (“Tc”). 图7和8中各子图的相关实验数据分别对应表3的(I)和(II)部分.

图7中的蒙板块与图8中白色矩形框外的蒙板块都代表虚检块, 而图8中白框内未出现蒙板的区域表明漏检区域, 其中虚检块的数量与保持内容的处理强度、篡改比例 (p_t) 及图像的纹理分布有关. 由于图像的平滑区域蕴含的信息较少, 因而经双重压缩处理后水印信息容易丢失, 所以在这些区域出现虚检的概率较高. 此外, 由于本方案采用了特征水印的交叉嵌入方法, 使得本文方案能抵制拼贴攻击, 检测效果及数据分别如图8(a)~(d)与表3(II)所示; 但同时也增加了在篡改区域外部分数据块的水印与篡改区内的对应块的特征不一致的概率, 因而出现图8的虚检块数总体高于图7的现象. 从图8中各子图的篡改区域容易看出, 白框内近一半的区域出现漏检现象, 这是由单个数据块嵌入1比特鲁棒性水印所致, 有关分析详见5.4小节最后一段.

依据于表3中的BER值(式(17))和TCR值(式(18)), 再根据4.3小节的检测策略, 可将图7和8中各子图判定(Auth.)为以下结果之一: 可信 (“Y”)、基本可信 (“Y?”)、不可信 (“N”)、不含水印 (“NW”). 基于表3最后一列数据的统计, 本次实验判定正确率为91.67% (22/24), 误判率为0%, 需要人工参与的判定, 即不确定性判定率为8.33% (2/24); 而这些性能指标与量化参数 θ_q , $\tau_1 \sim \tau_3$ 的设置有关, 最终取决于应用环境.

此外, 表3就第5.4小节涉及到的一些篡改检测指标如鲁棒性 (P_{t-Tr})、定位虚警率 (P_{t-lfp})、定位漏警率 (P_{t-lfn}), 分别给出了相应的理论值(Theory)与实验值(Experiment); 由于定位的准确率 (P_{t-ld}) 是根据 P_{t-lfp} 与 P_{t-lfn} 推导而出的(式(38)), 考虑到版面布局, 相关数据在表3中未列出. 表3(I)的最后一行显示, 在保持内容处理情况下, P_{t-Tr} 与 P_{t-lfp} 的理论与实验值的平均偏差, 即 $\text{Avg}(|\Delta|)$, 分别为

图 7 (网络版彩图) 双重压缩下保持内容处理后的检测蒙板图: 子图题代表图像经历的信号处理, DC_i 为第 i 种双重压缩方案, A_i 为两次压缩期间的攻击编号

Figure 7 (Color online) Detection mask images after content authentication for the content-preserving image attack in the double compression environment. Subtitles show the detailed process of signal processing; “ DC_i ” stands for i th-type double compression scheme; “ A_i ” is the number of attacks in Subsection 2.4. (a) Jpeg(75); (b) Jp2k(8); (c) DC_1 ; (d) DC_2 ; (e) DC_3 ; (f) DC_4 ; (g) DC_5 ; (h) $DC_1 \leftarrow A_3$; (i) $DC_2 \leftarrow A_6$; (j) $DC_3 \leftarrow A_8$; (k) $DC_4 \leftarrow A_{11}$; (l) $DC_5 \leftarrow A_2$

图 8 (网络版彩图) 双重压缩下恶意篡改后的检测蒙板图: 子图题代表图像经历的信号处理, DC_i 为第 i 种双重压缩方案, A_i 为两次压缩期间的攻击编号, T 与 T_c 分别为双重压缩期间的内容篡改攻击与拼贴攻击

Figure 8 (Color online) Detection mask images after content authentication for malicious content tampering attack in the double compression environment. Subtitles show the detailed process of signal processing; “ DC_i ” stands for i th-type double compression scheme; “ A_i ” is the number of attacks in Subsection 2.4; “ T ” and “ T_c ” respectively represent content tampering and collage attack. (a) Jpeg(75) T_cA_1 ; (b) Jp2k(8) T_c ; (c) $DC_1 \leftarrow T_c$; (d) $DC_2 \leftarrow T_c$; (e) $DC_3 \leftarrow T$; (f) $DC_4 \leftarrow T$; (g) $DC_5 \leftarrow T$; (h) $DC_1 \leftarrow TA_3$; (i) $DC_2 \leftarrow TA_6$; (j) $DC_3 \leftarrow TA_8$; (k) $DC_4 \leftarrow TA_{11}$; (l) $DC_5 \leftarrow TA_2$

表 3 双重压缩环境下的篡改检测结果和理论与实验数据的比较 ($\tau_1 = 0.385$, $\tau_2 = 0.263$, $\tau_3 = 0.333$)
Table 3 Tampering detection results and comparison of theoretical and experimental data under double compression

	Sub-graph	p_t (%)	P_{t_Tr}		P_{t_fp}		P_{t_fn}		BER	TCR	Auth.
			Theory	Experiment	Theory	Experiment	Theory	Experiment			
Figure 7	(a)			1.000		0.058			0.058	0.000	Y
	(b)			0.982		0.127			0.127	0.139	Y
	(c)			0.975		0.195			0.195	0.130	Y
	(d)			0.971		0.170			0.170	0.172	Y
	(e)			0.998		0.037			0.037	0.053	Y
	(f)			0.952		0.180			0.182	0.266	Y?
	(g)	0.0	0.977	0.990	0.023	0.052	0.0	0.0	0.052	0.189	Y
	(h)			0.987		0.169			0.169	0.075	Y
	(i)			0.913		0.227			0.227	0.313	Y?
	(j)			1.000		0.046			0.046	0.000	Y
	(k)			0.969		0.175			0.175	0.179	Y
	(l)			0.957		0.218			0.218	0.197	Y
Avg($\cdot$)			$ \Delta = 0.002$		$ \Delta = 0.119$		$ \Delta = 0.0$		0.142		
Figure 8	(a)	10.55	0.939	0.917	0.036	0.031	0.051	0.054	0.226	0.368	N
	(b)	16.41	0.913	0.894	0.043	0.050	0.076	0.107	0.259	0.411	N
	(c)	20.31	0.890	0.894	0.049	0.041	0.096	0.138	0.283	0.376	N
	(d)	24.90	0.857	0.868	0.057	0.067	0.121	0.185	0.296	0.446	N
	(e)	11.82	0.935	0.943	0.037	0.008	0.056	0.069	0.146	0.389	N
	(f)	17.19	0.908	0.870	0.044	0.062	0.081	0.104	0.278	0.467	N
	(g)	22.85	0.870	0.885	0.054	0.010	0.111	0.123	0.243	0.474	N
	(h)	27.34	0.843	0.837	0.060	0.043	0.131	0.153	0.333	0.490	N
	(i)	7.91	0.953	0.899	0.032	0.075	0.035	0.054	0.301	0.334	N
	(j)	29.30	0.829	0.891	0.063	0.003	0.141	0.187	0.268	0.409	N
	(k)	40.43	0.749	0.758	0.078	0.060	0.197	0.222	0.380	0.638	N
	(l)	50.78	0.671	0.649	0.090	0.069	0.253	0.227	0.442	0.792	NW
Avg($\cdot$)			$ \Delta = 0.023$		$ \Delta = 0.023$		$ \Delta = 0.027$				

0.002, 0.119; BER 均值为 0.142, 也就是 BCR 的均值为 0.858 (1-BER), 这与式 (23) 得到的 BCR 理论值 0.792 偏差 0.066; 再由表 3(II) 的最后一行可以看出, 在恶意篡改情况下 P_{t_Tr} , P_{t_fp} , P_{t_fn} 的对应偏差分别为 0.023, 0.023, 0.027. 这些偏差的存在与测试样本数量、样本个体差异, 以及信号处理强度有关. 总得来说, 上述相关测试指标的理论值与实验值之间偏差较小, 说明在 2.4 小节与第 3 节测试数据基础上形成的相关推导 (第 5 节), 对方案的实际应用具有较好的指导价值.

6.4 综合性能比较

在前述实验的基础上, 本小节就相关水印方案的性能做个综合比较. 用于水印技术应用的评价指

表 4 方案性能综合比较^{a)}

Table 4 Comprehensive comparison for related schemes

Scheme	Transparency		Robustness			Security		Watermark detection		Capacity (block size)
	PSNR	SSIM	Uc. (BCR)	Dc. (BCR)	Ds. (Std.)	Targeted attack	Collage attack	τ_1 (BER)	$P_{\tau_1, \text{FP}}$	
Ours	41.07	0.9897	0.948	0.854	0.0311	Y	Y	0.385	6.84×10^{-14}	16
Qi	40.3	0.9701	0.73	0.623	0.1092	N	N	0.484	10^{-6}	4
Wang	34	0.9627	0.85	0.733	0.0249	N	N	—	—	8
Hsu	38.19	0.965	0.913	0.867	0.0803	N	N	—	—	8

a) The bold font represents the optimal value of the corresponding column.

“—” means the content is not mentioned in the scheme.

标较多,而不同的应用领域采用的评价指标往往不同;根据内容认证领域对水印的常用需求,我们列举的评价指标如表 4 所示.其中 PSNR, SSIM 数据来自表 1 和 2,鲁棒性评价中的无压缩 (Uc.) 与双重压缩 (Dc.) 情形的数据分别来自图 6(e) 和 (f),而分布稳定性 (Ds.) 是指鲁棒性在不同攻击下的波动情况,我们基于图 6(f) 数据采用标准差来度量.有关安全性中的抗水印针对性攻击 (targeted attack) 与拼贴攻击 (collage attack),这完全取决于具体的水印算法 (参考第 1 节第 6 段).

由于水印在容量、鲁棒性、透明性上相互制约,要寻求一个方案在所有性能指标上都优于已有的其他方案是不现实的,表 4 中的数据也体现了这点.综合表 4 中的各项数据,本文方案综合性能优于其他 3 种方案.此外,本文的目标是要在双重压缩环境下实现内容认证,因此水印在透明性、容量与安全性符合实用的前提下,更应突出其抗攻击的鲁棒性与普适性,从这个角度比较,本文方案具有明显的优势.

7 结论

本文基于旋转向量的方法,提出了一种新颖的水印嵌入算法,并在此基础上构建了内容认证方案.理论分析与实验结果表明,本文的水印表达能在一定程度上抓住载体的语义,从而使我们的方案能容忍双图像重压缩.与相关半脆弱水印方案相比,本文方案具有以下特点:(1) 水印嵌入算法新颖,且有理论基础,可引入到相关领域开展应用研究;(2) 鲁棒性好,适用于双重压缩环境;(3) 在双重压缩环境下针对多种图像处理攻击,鲁棒性分布稳定,推广了方案的应用范围;(4) 安全性好,能同时抵制水印针对性攻击、拼贴攻击与伪造攻击.在双重压缩环境下,如何有效降低篡改定位的虚警率和漏警率,值得进一步研究.

参考文献

- 1 Lin C Y, Chang S F. Semi-fragile watermarking for authenticating JPEG visual content. In: Proceedings of SPIE, Security and Watermarking of Multimedia Contents II, San Jose, 2000. 140–151
- 2 Chang C C, Tsai P, Lin C C. SVD-based digital image watermarking scheme. Pattern Recogn Lett, 2005, 26: 1577–1586
- 3 Li C, Huang J W. A semi-fragile image watermarking resisting to JPEG. J Softw, 2006, 17: 315–324 [李春, 黄继武. 一种抗 JPEG 压缩的半脆弱图像水印算法. 软件学报, 2006, 17: 315–324]
- 4 Gong D F, Liu F L, Luo X Y. An image edge based robust watermarking algorithm. Sci Sin Inform, 2013, 43: 1410–1430 [巩道福, 刘粉林, 罗向阳. 一种基于图像边缘的鲁棒水印算法. 中国科学: 信息科学, 2013, 43: 1410–1430]
- 5 Ma L, Zhang X H. Characteristics of color images with watermark based on the relationship between non-void subspaces of inner space. Chinese J Comput, 2017, 40: 1203–1217 [马玲, 张晓辉. 基于内积空间非空子空间变换关系的含水印

- 彩色图像特征分析. 计算机学报, 2017, 40: 1203–1217]
- 6 Maeno K, Sun Q B, Chang S F, et al. New semi-fragile image authentication watermarking techniques using random bias and nonuniform quantization. *IEEE Trans Multimedia*, 2006, 8: 32–45
 - 7 Patra J C, Phua J E, Bornand C. A novel DCT domain CRT-based watermarking scheme for image authentication surviving JPEG compression. *Digit Signal Process*, 2010, 20: 1597–1611
 - 8 Zhang X, Cui L, Shao L. A fast semi-fragile watermarking scheme based on quantizing the weighted mean of integer Haar wavelet coefficients. In: *Proceedings of Symposium on Photonics and Optoelectronics*, Shanghai, 2012
 - 9 Preda R O. Semi-fragile watermarking for image authentication with sensitive tamper localization in the wavelet domain. *Measurement*, 2013, 46: 367–373
 - 10 Das C, Panigrahi S, Sharma V K, et al. A novel blind robust image watermarking in DCT domain using inter-block coefficient correlation. *AEU-Int J Electron Commun*, 2014, 68: 244–253
 - 11 Al-Haj A, Amer A. Secured telemedicine using region-based watermarking with tamper localization. *J Digit Imag*, 2014, 27: 737–750
 - 12 Guo J T, Zheng P J, Huang J W. Secure watermarking scheme against watermark attacks in the encrypted domain. *J Visual Commun Image Represent*, 2015, 30: 125–135
 - 13 Qi X J, Xin X. A singular-value-based semi-fragile watermarking scheme for image content authentication with tamper localization. *J Visual Commun Image Represent*, 2015, 30: 312–327
 - 14 Chen S T, Huang H N, Kung W M, et al. Optimization-based image watermarking with integrated quantization embedding in the wavelet-domain. *Multimed Tools Appl*, 2016, 75: 5493–5511
 - 15 Imran S, Kumar D P, Tetsuya S. A semi-fragile watermarking method using slant transform and LU decomposition for image authentication. In: *Proceedings of International Conference on Electrical, Computer and Communication Engineering*, Cox's Bazar, 2017. 881–885
 - 16 Wang Y, Pearmain A. Blind image data hiding based on self reference. *Pattern Recogn Lett*, 2004, 25: 1681–1689
 - 17 Hsu L Y, Hu H T. Robust blind image watermarking using crisscross inter-block prediction in the DCT domain. *J Visual Commun Image Represent*, 2017, 46: 33–47
 - 18 Parah S A, Sheikh J A, Loan N A, et al. Utilizing neighborhood coefficient correlation: a new image watermarking technique robust to singular and hybrid attacks. *Multidim Syst Sign Process*, 2018, 29: 1095–1117
 - 19 Ali M, Ahn C W. An optimal image watermarking approach through cuckoo search algorithm in wavelet domain. *Int J Syst Assur Eng Manag*, 2018, 9: 602–611
 - 20 Verma V S, Jha R K, Ojha A. Digital watermark extraction using support vector machine with principal component analysis based feature reduction. *J Visual Commun Image Represent*, 2015, 31: 75–85
 - 21 Mehta R, Rajpal N, Vishwakarma V P. A robust and efficient image watermarking scheme based on Lagrangian SVR and lifting wavelet transform. *Int J Mach Learn Cyber*, 2017, 8: 379–395
 - 22 Holliman M, Memon N. Counterfeiting attacks on oblivious block-wise independent invisible watermarking schemes. *IEEE Trans Image Process*, 2000, 9: 432–441
 - 23 Goljan M. Cryptanalysis of the Yeung-Mintzer fragile watermarking technique. *J Electron Imag*, 2002, 11: 262–274
 - 24 Wang Z, Bovik A C, Sheikh H R, et al. Image quality assessment: from error visibility to structural similarity. *IEEE Trans Image Process*, 2004, 13: 600–612

A watermarking algorithm for image content authentication in double-compression environment

Jianjing FU^{1,2*}, Deren CHEN¹, Dawen XU³ & Jiafa MAO⁴

1. College of Computer Science and Technology, Zhejiang University, Hangzhou 310058, China;

2. College of New Media, Communication University of Zhejiang, Hangzhou 310018, China;

3. School of Electronics and Information Engineering, Ningbo University of Technology, Ningbo 315016, China;

4. College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou 310023, China

* Corresponding author. E-mail: fjjsn@163.com

Abstract Malicious tampering of digital image content may lead to serious consequences. As image resolution increases and its circulations on the network grows rapidly, images undergo double compression in the following environment: compression (JPEG/JPEG2000) for first release, decoding, application processing (conventional signal processing/security attacks/malicious content tampering), and compression (JPEG2000/JPEG) for rerelease before detection. Effectively determining whether the image content has been tampered and to locate the tampering position in a double-compression environment is an urgent problem. In this paper, a novel watermark expression method based on rotating vector and its modulation algorithm is proposed. Based on this method, a semi-fragile watermarking scheme is established for the image content authentication. The stability of watermarked data is analyzed theoretically, providing the watermark scheme with a theoretical basis. The authentication schemes with feature extraction and reconstruction, watermark embedding and extraction, and tamper detection and localization are mainly elaborated. The robustness of watermark, its security, and related detection performance are analyzed. Theoretical analysis and experiments show that the scheme features a good watermark transparency, good robustness, and stable distribution for different attacks and can effectively distinguish malicious tampering from content-preserving operators and locate tampering area under double-compression environment. In addition, the proposed scheme can resist watermark attacks, collage attacks, and forge attacks. Compared with related schemes, the proposed scheme has superior comprehensive performance and is suitable for content authentication in double-compression environment, which expands the application scope of watermark-based content authentication.

Keywords image watermark, semi-fragile watermark, malicious tampering, content authentication, double compression

Jianjing FU received his M.S. degree from Department of Computer Science, Kunming University of Science and Technology, China, in 2002, and Ph.D. degree from Institute of Remote Sensing and Information Technology, Zhejiang University, Hangzhou, China, in 2012. From 2014 to 2016, he was a post-doc researcher in College of Computer Science and Technology, Zhejiang University. Since 2012, he has been with Communication University of Zhejiang, Hangzhou, where he has been an associate professor with College of New Media. His research interests include software protection, information hiding and digital watermarking, and digital content security.

Deren CHEN was born in 1951, and received his B.E. degree in applied mathematics and M.S. degree in computer application from Zhejiang University, Hangzhou, China. He has been a professor and doctoral tutor at School of Computer Science and Technology, Zhejiang University, associate dean at School of Software, Zhejiang University, dean of School of Software of Zhejiang University (Ningbo), and director at Electronic Service Research Center of Zhejiang University. In 2000, he was a visiting scholar of IBM T. J. Watson Research Center, New York, and a visiting professor in School of Applied Science, Nanyang Technological University, Singapore, in 1999. His research interest includes electronic commerce, data mining, and information security.

ing in the encrypted domain.

Dawen XU received his M.S. degree in communication and information system from Ningbo University, China, in 2005 and Ph.D degree in computer applied technology from Tongji University, China, in 2011. He is a professor at School of Electronics and Information Engineering, Ningbo University of Technology, China. His research interests focus on digital watermarking and information hiding and signal processing.

interests include pattern recognition, digital image processing, and information hiding.

Jiafa MAO received his Ph.D. degree in pattern recognition from East China University of Science and Technology, China, in 2009. Since then, he has worked as a post-doc researcher at Beijing University of Posts and Telecommunications, China. In July 2011, he joined Zhejiang University of Technology, China, where he is currently a full time professor in School of Computer Science and Technology. His research