◎网络、通信与安全◎

基于超混沌的医学图像篡改定位零水印算法

肖振久[1,2],李 南[1],王永滨[2],姜正涛[2],陈 虹[1]
XIAO Zhenjiu[1,2], LI Nan[1], WANG Yongbin[2], JIANG Zhengtao[2], CHEN Hong[1]

1.辽宁工程技术大学 软件学院,辽宁 葫芦岛 125100
2.中国传媒大学 计算机学院,北京 100024
1. Software College, Liaoning Technical University, Huludao, Liaoning 125100, China
2. School of Computer Science, Communication University of China, Beijing 100024, China

XIAO Zhenjiu, LI Nan, WANG Yongbin, et al. Zero watermarking scheme for medical image temper location based on hyper-chaos encryption. Computer Engineering and Applications, 2017, 53(7):115-120.

Abstract: Medical image facing malicious tampering, information thefting and leaking, this paper proposes a hyper-chaos based zero watermarking scheme for image authentication and tamper detection. Firstly, original carrier image's LSB plane should be initialized to zero. Secondly it is divided into non-overlapping blocks and calculating the mean value of each block, using the size relationship between each pixel of block and mean value of each block to construct the characteristic matrix. Finally, Arnold scrambled characteristics matrix and hyper-chaotic encrypted binary watermark are executed for XOR operation to construct a zero-watermark. In addition, if the image has been tampered, tampered area can be detected and localized accuracy by calculating the difference image. The experimental results show that the scheme not only has good security, but also achieves superior tamper detection and localization.
Key words: hyper-chaos encryption; tamper detection; zero watermarking; medical image

摘 要:针对医学图像面临恶意篡改、信息窃取和泄露等安全问题,提出了一种基于超混沌的医学图像篡改定位零水印算法。首先将原始载体图像的最低有效位LSB(Least Singificant Bit)置零,然后将其分成互不重叠的子块,计算每块的均值,利用块中每个像素值与块均值的大小关系构造特征矩阵,最后利用Arnold置乱后的特征矩阵与超混沌加密的二值水印进行异或运算来构造零水印;另外,如果图像遭到篡改,通过计算差值图像,则能精确定位篡改位置及篡改形状。实验结果表明,该方案不仅具有较好的安全性,而且达到良好的篡改检测和定位效果。
关键词:超混沌;篡改检测;零水印;医学图像
文献标志码:A **中图分类号:**TP391 doi:10.3778/j.issn.1002-8331.1510-0128

1 引言

伴随现代通信技术和生物医学工程的快速发展,医院与医院之间的远程医疗信息交换变得更快、更容易。但是在传输过程中,医学数据有可能被有意或无意地操纵,这可能会导致严重的后果。由于医学图像包含了一些重要的病理信息,任何较小的改动,都会影响医生对病理的判断。如何确保医学图像的完整性、安全性是现代医学领域面临的一大难题。

数字水印尤其是脆弱水印[1-3]和半脆弱水印[4-6]的出现为原始图像的认证和恶意篡改区域的检测及恢复提

基金项目:国家自然科学基金(No.61103199);教育部-中移动科研基金(No.MCM20130411)。
作者简介:肖振久(1968—),男,副教授,主要研究方向:网络与信息安全、数字版权管理;李南(1991—),女,硕士研究生,主要研究方向:网络信息安全与数字水印,E-mail:lisarylong@163.com;王永滨(1963—),男,教授,研究方向为:形式化建模与仿真、分布式计算等;姜正涛(1976—),男,副教授,主要研究方向为:密码学与信息安全;陈虹(1967—),女,副教授,主要研究方向:网络安全。
收稿日期:2015-10-15 **修回日期:**2015-12-01 **文章编号:**1002-8331(2017)07-0115-06
CNKI**网络优先出版:**2015-12-23, http://www.cnki.net/kcms/detail/11.2127.TP.20151223.1505.002.html

供了最佳解决方案。脆弱水印和半脆弱水印对于微小的篡改极为敏感,可以对篡改区域进行精确定位,因而得到广泛关注。

文献[7-10]提出了一种基于混沌的脆弱水印算法,该算法通过修改载体图像的最低有效位来实现水印的嵌入。虽然该算法具有篡改区域定位能力,但被篡改区域有可能与嵌入信息相匹配,篡改区域不能完全被确定。另外,将水印信息添加到载体图像中,无法保证原始图像数据不被干扰,无法满足医学领域对图像质量非常严格的要求。文献[11]提出了一种新颖的图像认证算法,该算法将医学图像划分为感兴趣区域和非感兴趣区域,将图像的感兴趣区域和电子病例进行联合稀疏编码嵌入到医学图像的非感兴趣区域。该算法不但实现了医学图像感兴趣区域的篡改定位,而且还具有修复篡改感兴趣区域的能力,在一定程度上减少了对原始图像的破坏,但都无法保障原始图像的完整性。

温泉[12]等人提出的零水印概念为解决该问题开辟了新的途径。该方法在不改变宿主图像的前提下,利用图像的特征信息和水印数据构造零水印,有效地解决了图像透明性和鲁棒性之间的矛盾。

文献[13]提出一种将Arnold置乱变换与DCT相结合的医学零水印算法,该算法具有很好的透明性和鲁棒性,但无法检测和定位图像的篡改区域。文献[14]提出的医学图像篡改定位零水印算法可以有效地解决该问题,该算法选取二值水印与置乱载体图像的最低有效位执行异或操作来构造零水印,实验证明该算法能够定位篡改位置及篡改形状,但难以抵抗噪声攻击。

本文提出了一种基于图像篡改的零水印方法。该方法在不改变医学图像原始信息的前提下,利用图像的特征信息和水印信息构造的零水印来完成图像的认证,同时通过篡改图像提取水印与原水印的差值来精确定位篡改的位置及形状,而且鲁棒性也得到一定的提升。

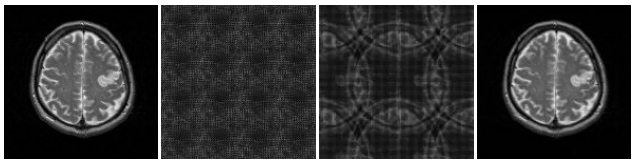
2 相关知识

2.1 Arnold置乱

Arnold变换是一种传统的混沌系统,又称“猫脸变换”。该变换通过像素空间位置变换达到“杂乱无章”的效果,减少了图像相邻像素间的关联性,提高了图像安全性。因此,本文选择Arnold变换对图像进行预处理。对于一幅 $N\times N$ 的二值图像,(x,y) 表示原图像像素坐标,(x',y') 表示变换后的像素坐标,N 为图像的阶数,变换形式如式(1)所示:

$$\begin{pmatrix}x'\\y'\end{pmatrix}=\begin{pmatrix}1&1\\1&2\end{pmatrix}\begin{pmatrix}x\\y\end{pmatrix}(\mathrm{mod}\,N) \tag{1}$$

该变换具有一定的周期性,即经过有限次的迭代便可恢复出原始图像。图1是大小 256×256 的CT图像不同迭代次数下的置乱图像,经过一个周期192次迭代后图像恢复到初始状态。本算法中,周期和迭代次数作为密钥 K_1 存在。



原始图像 48次置乱 96次置乱 192次置乱

图1 Arnold置乱图

2.2 混沌序列与二维超混沌离散系统

混沌系统[15-16]具有遍历性、初值敏感性和伪随机性等良好特性,非常适合应用到图像加密领域。

目前被广泛研究的Logistic映射是一种形式简单的一维混沌系统。该混沌系统对初值极为敏感,所生成的混沌序列具有非周期性和不收敛性,其一维差分形式的映射方程为:

$$x_{i+1}=\mu x_i(1-x_i) \tag{2}$$

当 $3.569\,945\,6<\mu\leqslant4$ 时,Logistic映射处于混沌状态。

低维混沌系统存在一定的缺陷,主要表现在动力行为相对比较简单、确定性序列的参数相对较少(密钥空间少)以及安全性能相对较低等方面。而高维超混沌系统则具有更多的方向不稳定性、更复杂的动力学行为以及更大的密钥空间。水印算法对水印安全性和随机性的要求能够被更好地满足[6,17]。故本文采用二维超混沌系统,其一般形式为:

$$\begin{cases}x_{n+1}=a_1+a_2x_n+a_3x_n^2+a_4y_n+a_5y_n^2+a_6x_ny_n\\y_{n+1}=a_7+a_8x_n+a_9x_n^2+a_{10}y_n+a_{11}y_n^2+a_{12}x_ny_n\end{cases} \tag{3}$$

式中 $a_i=(1,2,\cdots,12)$ 均为待定常系数。表1为一些形式简单且具有超混沌特性的二维离散系统[17]。

表1 形式简单的二维超混沌离散系统

系统编号	二维离散方程	参数设定	Lyapunov指数
1	$\begin{cases}x_{n+1}=a_4y_n+a_5y_n^2\\y_{n+1}=a_8x_n+a_{10}y_n\end{cases}$	$a_4=1.55$ $a_5=-1.3$ $a_8=-1.1$ $a_{10}=0.1$	0.238 0.166
2	$\begin{cases}x_{n+1}=a_5y_n^2\\y_{n+1}=a_7+a_9x_n^2+a_{10}y_n\end{cases}$	$a_5=1.3$ $a_7=-1.05$ $a_8=1.15$ $a_{10}=-0.2$	0.211 0.046
3	$\begin{cases}x_{n+1}=a_2x_n+a_4y_n\\y_{n+1}=a_7+a_9x_n^2+a_{10}y_n\end{cases}$	$a_2=-0.95$ $a_4=1.55$ $a_7=-0.45$ $a_9=2.4$ $a_{10}=1.05$	0.302 0.240
4	$\begin{cases}x_{n+1}=a_4y_n+a_6x_ny_n\\y_{n+1}=a_7+a_8x_n+a_{10}y_n\end{cases}$	$a_4=-0.95$ $a_6=-1.1$ $a_7=0.55$ $a_8=1.55$ $a_{10}=-1.8$	0.175 0.066

选择表1中系统编号为1的二维离散方程对水印进行加密，将各参数代入其中，其结果如式(4)所示，当 $a_4=1.55$，系统进入超混沌状态。

$$\begin{cases}x_{n+1}=1.55y_n-1.3y_n^2\\y_{n+1}=-1.1x_n+0.1y_n\end{cases} \tag{4}$$

3　基于超混沌的医学图像零水印算法

将Arnold置乱、超混沌离散系统和最低有效位LSB算法巧妙结合，提出了一种具有篡改定位能力的零水印算法。本文算法首先应用超混沌序列对水印图像进行加密预处理；然后将最低有效位置零的载体图像划分成互不重叠的子块，计算子块的均值，利用块中每个像素与块均值的大小关系来构造特征矩阵；最后利用Arnold置乱后的特征矩阵与超混沌加密后的二值水印进行异或运算来实现零水印的嵌入。水印提取过程与嵌入过程非常相似，将提取水印与原水印作绝对差值运算即可得到图像的篡改区域。

算法利用超混沌离散系统对水印信息进行预处理，不但增强了水印的安全性，而且提高了水印信号在载体图像中的隐蔽性，另外零水印的嵌入也不会影响原始载体图像的质量。因此本文算法不但具有较好的透明性，而且表现出较强的鲁棒性，同时也能够对图像篡改区域进行检测和定位。

3.1　水印图像超混沌加密

本文算法使用 $N\times N$ 的二值水印图像 $W(i,j)$，其中 $0<i\leqslant N$，$0<j\leqslant N$。

将初始值 x_0、y_0 作为密钥 K_2，应用式(4)产生长度为 $N\times N$ 的二维超混沌序列 $x(k)$ 和 $y(k)$，其中 $k=1,2,\cdots,N\times N$。为保证其超混沌特性，采用式(5)的参数降维模型将二维序列 $x(k)$ 和 $y(k)$ 转换成一维序列 $L(k)$，该序列具有较为理想的扩散均匀度[6]：

$$L(k)=\frac{[x(k)-y(k)+1.5]}{2.5} \tag{5}$$

图2为该模型下产生的 256×256 点降维后的超混沌序列。

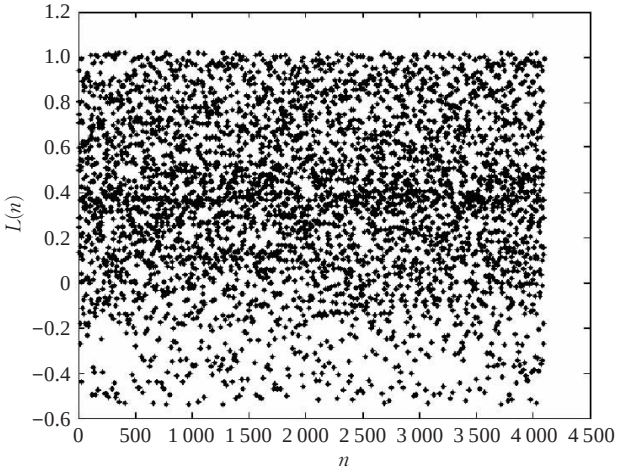


图2　二维超混沌序列经降维后的序列

系统加密设计如下：

首先将一维序列 $L(k)$ 转化为二维矩阵 $LL(i,j)$，然后按式(6)、(7)对水印进行加密处理。

$$C(i,j)=W(i,j)\oplus J(i,j) \tag{6}$$

$$J(i,j)=\begin{cases}1, & LL(i,j)>ml\\0, & LL(i,j)\leqslant ml\end{cases},\ ml=mean(LL) \tag{7}$$

其中 $J(i,j)$ 为二值化的超混沌序列，ml 为二维矩阵 LL 的均值，$C(i,j)$ 为经超混沌加密后的二维水印数组。

系统解密设计如下：

$$W(i,j)=C(i,j)\oplus J(i,j) \tag{8}$$

3.2　零水印的构造

设原始载体CT图像 I 大小为 $M\times M$，零水印的构造过程如下：

(1)将原始载体图像I的最低有效位LSB置零，得到图像 I_LS_0。

(2)将 I_LS_0 划分成互不重叠的大小为 $m\times m$ 的子块，每个子块记为 P_k（$k=1,2,\cdots,\frac{M\times M}{m\times m}$）。

(3)计算每个子块 P_k 的平均值 T_k：

$$T_k=\frac{1}{m^2}\sum_{i=1}^{m}\sum_{i=1}^{m}P_k(i,j),\ i,j=1,2,\cdots,m \tag{9}$$

(4)通过对比块中每个像素值与块均值的大小关系来构造一个大小为 $M\times M$ 的二值矩阵 B，其每块像素值为 b_k。

$$b_k(i,j)=\begin{cases}1, & P_k(i,j)>T_k\\0, & P_k(i,j)\leqslant T_k\end{cases} \tag{10}$$

(5)对二值矩阵 B 作 k 次Arnold变换，得到置乱后的图像 B_s。

(6)将加密后的水印序列 C 与 B_s 进行异或运算，得到长度为 $N\times N$ 的认证序列 LS_1，该序列用于水印提取和定位篡改区域。

零水印的构造过程如图3所示。

3.3　水印检测及篡改定位过程

水印检测及篡改定位过程如下：

(1)将待检测图像 I' 的最低有效位置零，得到 I'_LS_0。

(2)把 I'_LS_0 划分成互不重叠的大小为 $m\times m$ 的子块，每个子块记为 P'_k（$k=1,2,\cdots,\frac{M\times M}{m\times m}$）。

(3)计算每个子块 P'_k 的平均值 T'_k：

$$T'_k=\frac{1}{m^2}\sum_{i=1}^{m}\sum_{i=1}^{m}P'_k(i,j),\ i,j=1,2,\cdots,m \tag{11}$$

(4)通过对比块中每个像素值与块均值的大小关系构造一个大小为 $M\times M$ 的二值矩阵 B'，其每块的像素值为 b'_k。

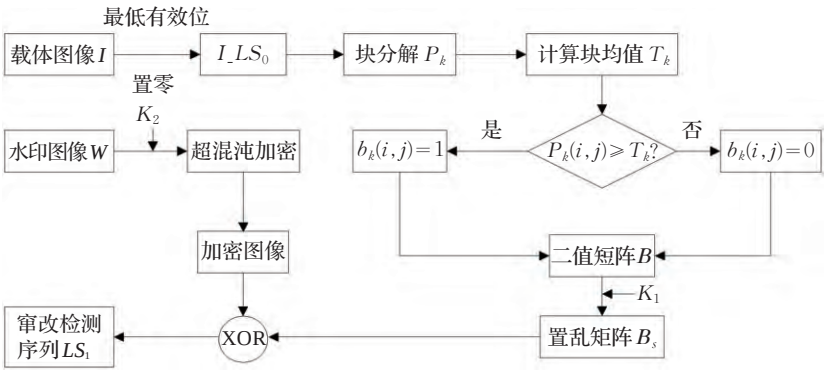


图3　零水印构造流程图

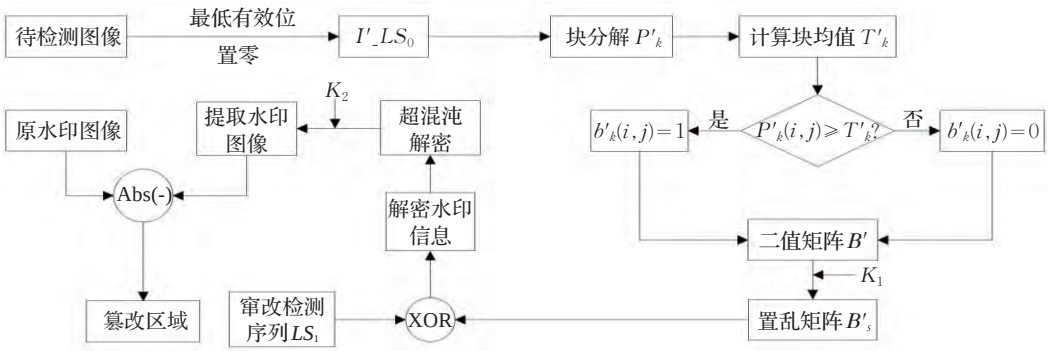


图4　水印检测及篡改定位流程图

$$b'_k(i,j)=\begin{cases}1, & P'_k(i,j)>T'_k\\ 0, & P'_k(i,j)\leqslant T'_k\end{cases} \tag{12}$$

(5)对二值矩阵 B' 作 k 次 Arnold 变换,得到置乱后的图像 B'_s。

(6)用水印嵌入时生成的序列 $LS1$ 与 B'_s 异或得到水印信息 C_1,将 C_1 按式(7)和式(8)进行超混沌解密,得到提取出的水印图像 W'。

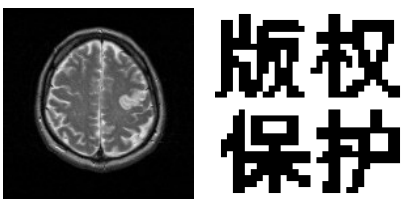
(7)计算差值图像 $|W'-W|$,即可得到图像被篡改的区域。

水印检测及篡改定位过程如图4所示。

4　仿真实验及结果分析

4.1　实验参数

实验环境为 Matlab R2010b,实验选取 256×256 的 CT 图像为原始载体图像,256×256 的二值图像“版权保护”为水印图像,分别如图5(a)、(b)所示。实验中,$m=8$,Arnold 置乱算法中迭代次数 $k=75$ 和置乱周期 $T=192$ 作为密钥 K_1,超混沌系统参数选取 $x_0=0.123$, $y_0=0.321$ 作为密钥 K_2。

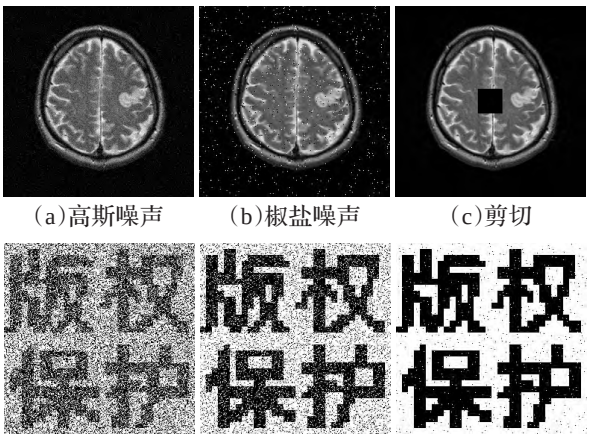


(a)原始CT图像　(b)水印图像

图5　原始载体图像和水印图像

4.2　算法的鲁棒性评价

为了检验算法的鲁棒性,分别对载体图像5(a)进行高斯噪声攻击(均值为0,方差为0.002)、椒盐噪声攻击(强度为0.02)以及剪切攻击(中心剪切1/8)。攻击后的载体图像和提取出的水印图像如图6所示。



(a)高斯噪声　(b)椒盐噪声　(c)剪切

(d)提取水印　(e)提取水印　(f)提取水印

图6　各种攻击后提取出的水印图像

本文采用归一化相关系数 NC 作为水印鲁棒性的客观评价标准,其定义如式(13)所示。NC 值越大,算法鲁棒性越强,反之亦然。

$$NC(x_1,x_2)=\frac{\sum_{i=1}^{M}\sum_{j=1}^{N}x_1(i,j)x_2(i,j)}{\sqrt{\sum_{i=1}^{M}\sum_{j=1}^{N}x_1(i,j)^2}\sqrt{\sum_{i=1}^{M}\sum_{j=1}^{N}x_2(i,j)^2}} \tag{13}$$

其中 x_1、x_2 分别表示原始水印和提取水印。

为了进一步说明本算法具有较好的鲁棒性,表2给出了图像受到攻击后提取水印与原水印图像的归一化相关系数,并与文献[14]中的算法对比。文献[14]利用图像最低有效位来构造特征矩阵,虽然是一种无损的零水印嵌入方式,但是LSB算法很容易被噪声和几何攻击等操作破坏,恢复出的水印图像与原始水印有一定的差别。本文利用LSB置零后分块像素值和块均值的大小关系构造特征矩阵,由于图像受到攻击时块均值的变化相对较小,具有很好的鲁棒性,而且零水印的嵌入是一种无损方式,因此,本文算法的水印提取效果更好。

表2　算法鲁棒性评价(*NC*值)

攻击类型	攻击强度	文献[14]	本文算法
高斯噪声	0.002	0.542 4	0.706 7
	0.004	0.539 0	0.683 8
椒盐噪声	0.020	0.869 2	0.885 7
	0.040	0.806 3	0.821 5
剪切	中心剪切1/8	0.932 7	0.989 5
	中心剪切1/4	0.901 5	0.963 2

4.3　篡改检测评价

为了验证本算法对于篡改检测的有效性,分别对载体图像实施剪切、文本添加和复制粘贴篡改攻击,其实验结果如图7~图9所示。

(a)剪切篡改　(b)提取水印　(c)篡改区域

图7　剪切篡改攻击

(a)添加文本　(b)提取水印　(c)篡改区域

图8　添加文本攻击

(a)复制粘贴　(b)提取水印　(c)篡改区域

图9　复制粘贴篡改攻击

图7为对载体CT图像进行剪切篡改后的检测效果。实验在CT图中裁剪了左上角的一块重要区域,因为剪切篡改区域相对图像整体来说比较小,所以从篡改后的载体中依然可以提取较为清晰的水印图像。另外算法成功定位了剪切的位置与形状,如图7(c)中显示的白色区域。

对载体图进行文本添加攻击,篡改后的检测效果如图8所示。在CT图像中心添加一行"CT"字样的文本,篡改图像如图8(a)所示,水印提取和篡改检测效果分别如图8(b)和图8(c)所示。从图8(c)中可以看出添加文本的位置和大概轮廓,表明算法对于本文添加攻击具有一定的检测能力。

图9中,在CT图像的中心域复制粘贴了载体图像右侧的一块白色区域,篡改图像如图9(a)所示,提取出的水印图像如图9(b)所示。可以看出提取水印和原始差别非常微小,而且成功对篡改区域进行了定位,如图9(c)所示。

从图7~图9看出本文算法对于复制粘贴、文本添加以及剪切篡改攻击均表现出良好的定位效果,而且篡改攻击后提取出的水印也较为清晰,表明算法在篡改检测方面具有较好的可行性。

为了更加客观地评价该算法对于篡改检测的有效性,表3给出了各种篡改攻击后图像的 *PSNR* 值以及提取水印的 *NC* 值。从表中可以看出篡改攻击后提取的水印 *NC* 值均在0.99之上,证明该算法对于篡改攻击表现出较好的鲁棒性。

表3　篡改攻击后图像的*PSNR*值和提取水印的*NC*值

篡改攻击类型	*PSNR*	*NC*
复制粘贴	33.251 8	0.993 4
文本添加	32.600 6	0.992 3
剪切	23.141 9	0.990 2

4.4　算法的透明性和安全性评价

由于零水印的嵌入没有对载体图像做任何更改,图像的透明性不会受到影响,攻击者无论使用何种方法也无法检测出水印的存在。超混沌系统和Arnold变换分别实现了水印信息和原始图像特征矩阵的二次加密,为算法的安全性提供了保证。虽然水印算法采用最简单的超混沌映射、Arnold变换和LSB 算法,但在密钥 K_1、K_2 不确定的情况下准确定位图像篡改位置也是相当困难的。

5　结束语

在医学研究中,保持医学图像的安全性和真实性是非常必要的。本文算法受到超混沌系统和Arnold变换的多密钥保护,提升了水印的安全性,同时与零水印的结合保证了图像的质量。另外该算法具有很好的鲁棒性,即使受到攻击者的恶意攻击,也可以清晰地还原出原始水印,实现了水印鲁棒性和透明性的统一。超混沌系统对初值的超强依赖性使得算法对图像修改表现出很好的脆弱性,可以有效地检测出篡改的位置及形状。

本文算法简单,易于实现,计算速度也快,具有一定的应用前景。今后的工作,将会把如何恢复图像篡改区域作为研究重点。

参考文献:

[1] 霍耀冉,和红杰,陈帆.基于邻域比较的JPEG脆弱水印算法及性能[J].软件学报,2012,23(9):2510-2521.

[2] Tong Xiaojun,Liu Yang,Zhang Miao,et al.A novel chaos-based fragile watermarking for image tampering detection and self-recovery[J].Signal Processing:Image Communication,2013,28(3):301-308.

[3] 张玉梅,和红杰,陈帆.浏览器端定位篡改的网页脆弱水印算法[J].计算机研究与发展,2014,51(12):2604-2613.

[4] 刘东彦,刘文波,张弓.图像内容可恢复的半脆弱水印技术研究[J].中国图象图形学报,2010,15(1):20-25.

[5] 赵春晖,刘巍.基于分块压缩感知的图像半脆弱零水印算法[J].自动化学报,2012,38(4):609-617.

[6] 杨晋霞,鞠杰,邵峰.基于超混沌加密的半脆弱音频水印算法[J].计算机应用与软件,2014,31(11):295-298.

[7] 蔡键,叶萍,刘涛.基于小波变换的用于医学图像的半脆弱水印算法[J].计算机应用与软,2011,28(6):278-230.

[8] 陈帆,王宏霞.定位像素篡改的安全脆弱水印算法[J].铁道学报,2011,33(1):63-68.

[9] Rawat S,Raman B.A chaotic system based fragile watermarking scheme for image tamper detection[J].AEU-International Journal of Electronics and Communications,2011,65(10):840-847.

[10] 刘敏,陈志刚,邓小鸿.基于混沌和脆弱水印的图像篡改检测算法[J].计算机应用,2013,33(5):1371-1373.

[11] Tareef A,Al-Ani A,Hung N,et al.A novel tamper detection-recovery and watermarking system for medical image authentication and EPR hiding[C]//Proceedings of the 36th Annual International Conference on Engineering in Medicine and Biology Society(EMBC),2014:5554-5557.

[12] 温泉,孙锬锋,王树勋.零水印的概念与应用[J].电子学报,2003,31(2):214-216.

[13] 隋淼,李京兵.一种基于Arnold置乱变换和DCT的医学图像鲁棒水印算法[J].计算机应用研究,2013,30(8):2552-2556.

[14] 吴伟民,丁冉,林志毅,等.基于混沌的医学图像篡改定位零水印[J].计算机应用研究,2014,31(12):3685-3688.

[15] Benrhouma O,Hermassi H,Ahmed A,et al.Chaotic watermark for blind forgery detection in images[DB/OL].[2015-09-30].http://link.springer.com.

[16] 刘瑶利,李京兵.一种基于DCT和Logistic Map的医学图像鲁棒多水印方法[J].计算机应用研究,2013,30(11):3430-3433.

[17] 周武杰,郁梅,禹思敏,等.一种基于超混沌系统的立体图像零水印算法[J].物理学报,2012,61(8):117-126.

(上接78页)

参考文献:

[1] 殷彬,杨会志.灵活结构网页的正文提取[J].计算机技术与发展,2011,21(9):111-113.

[2] Cai Deng,Yu Shipeng,Wen Jirong,et al.VIPS:A vision based on page segmentation algorithm,Microsoft Co,Tech Rep:MSR-TR-2003-79[R].2003.

[3] 韩忠明,李文正,莫倩,等.有效HTML文本信息抽取方法的研究[J].计算机应用研究,2008,25(12):3568-3571.

[4] 安增文,徐杰锋.基于视觉特征的网页正文提取方法研究[J].微型机与应用,2010,9(3):38-41.

[5] Ji Xiangwen,Zeng Jianping,Zhang Shiyong,et al.Tag tree template for Web information and schema extraction[J].Expert Systems with Applications,2010,37(12):8492-8498.

[6] 王少康,董科军,阎保平.使用特征文本密度的网页正文提取[J].计算机工程与应用,2010,46(20):1-3.

[7] 刘军,张净.基于DOM的网页主题信息抽取[J].计算机应用与软件,2010,27(5):188-190.

[8] Mantratzis G C,Orgun M A,Cassidy S.Separating XHTML content from navigation clutter using dom-structure block analysis[C]//Proceedings of Conference on Hypertext,2005:145-147.

[9] 杨钦,杨沐昀.一种基于标点密度的网页正文提取方法[J].智能计算机与应用,2015,5(4):42-44.

[10] 高屹.基于树先剪枝的网页正文抽取方法研究[J].科技创新与应用,2013(36):63-64.

[11] 张瑞雪,宋明秋,公衍磊.逆序解析DOM树及网页正文信息提取[J].计算机科学,2011,38(4):213-215.

[12] 朱逢春.基于DOM树的网页去噪技术[J].电子制作,2015(8).

[13] Liu W,Huang G,Liu X.Detection of pubishing Web pages based on visual similarity[C]//Proceedings of the 16th Intl Conf on World Wide Web,2007:61-70.

[14] Kai S,Lausen G.VIPER:Augmenting automatic information extraction with visual perceptions[C]//Proceedings of the ACM CIKM Int'l CorIf on Information and Knowledge Management.[S.l.]:ACM Press,2005:381-388.

[15] Chibane I,Doan B L.A Web page topic segmentation algorithm based on visual criteria and content layout[C]//Proceedings of SIGIR Coference,2007.