数据隐藏中图像去噪攻击及其对策

袁 渊 窦文华
(国防科技大学计算机学院 长沙410073)

摘 要 数字水印是数字版权保护中的关键技术,本文从水印攻击角度分析了去噪攻击的原理,并且提出了相应的解决办法,通过实验证明了理论分析的正确性,这对水印算法研究和设计有一定的参考价值。

关键词 数字水印,去噪攻击,最大后验概率估计,极大似然估计

Countermeasures to Image Denoising Attacks in Data Hiding

YUAN Yuan DOU Wen-Hua
(Computer College, National University of Defense Technology, Changsha 410073)

Abstract Digital watermark is a key technique of digital copyright protection. This paper analyzes the principle of denoising from the watermark attack point of view and presents several solutions to this problem. Experimental results prove the correctness of theory analysis. This would be of benefit to the design and research of watermark algorithm.

Keywords Digital watermark, Denoising attack, Maximum a posterior, Maximum likelihood

1. 引言

由于互联网和个人电脑的普及,人们能够方便地使用和分发数字多媒体信息,并且原有的模拟声音和视频设备也正在被后续的数字设备所替代。尽管数字信息和模拟信息相比,有许多优点,但是由于缺乏有效的保护数字作品版权的有效措施,有版权的数字作品被毫无限制地使用、复制和分发,导致服务提供商不愿意以数字媒体形式提供服务。为此,研究者开发了两种技术来保护数字作品的版权问题:密码学(Cryptography)和数字水印(digital watermark)。加密技术在数字信息从发送方到接收方传输过程中对数字信息加密,达到保护数字版权的目的。但是当接收者将接收到数据解密后,加密的数字信息就成为不受保护的信息。数字水印技术作为加密技术的补充,通过在数字作品中嵌入用户不可见的版权标识信息,如作品的作者、所有者、发行者和授权消费者等,这些信息只有在影响作品质量的情况下才能被删除,从而达到保护数字版权的目的。因此数字水印是目前进行数字作品版权保护的一种比较有效的技术手段。

为了保护数字作品版权,有效的数字水印方案必须具有不可感知性、鲁棒性和信息量。不可感知性是指嵌入的数字水印不应该影响数字作品视觉或听觉上的欣赏效果,与水印系统的应用和目的无关;鲁棒性是指嵌入水印后的数字作品被有意或恶意地处理和修改后,仍能够检测出水印信息,通常鲁棒性的等级随不同应用的要求而不同;信息量则要求嵌入的水印必须携带一定数量的信息,这些信息可以用来判定和跟踪数字作品的所有者和购买者。目前盗版已成为数字化产业最大的威胁,既使采用了数字水印技术,数字水印系统也必然面临着盗版者各式各样的攻击,通过研究各种攻击手段的原理和对策,能够指导数字水印系统研究和开发者的工作,从而设计出性能优良的数字水印系统。

2. 数字水印的通信模型及攻击手段

由于数字水印系统和数字通信系统的模型非常相似,因而可以借助数字通信系统中的概念和方法来描述数字水印系统的问题。通常把数字水印问题建模为扩频数字通信问题:在嵌入水印 w 的时候,原始图像 x 作为通信信道,信道传递概率为 $P_x(I/W)$,水印是要传送的消息;信息在传输过程中可能受到各种有意或无意的干扰,这些干扰构成了传输过程中的攻击信道,信道传递概率为 $P_a(I^*/W)$;信息到达接收方后,进行水印检测,从数字通信的角度来看,水印的检测也就是如何有效消除原始图像信道和攻击信道带来的噪声,见图1。

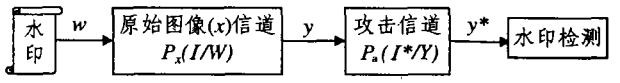


图1 数字水印通信模型

只考虑加性信息隐藏和加性攻击信道噪声,模型简化为:

$$y^* = w + s + a \quad (1)$$

其中 s 和 a 分别是原始图像信道和攻击信道引入的噪声。可以看出,对于私有水印方案,检测过程中需要原始图像,因而原始图像的噪声 s 可以从含水印图像中完全消除。对于公开水印系统,系统性能的优劣取决于如何有效地去除原始图像和各种攻击引入的噪声。

Craver 等人把水印攻击分为四类:鲁棒性攻击、表示攻击、解释攻击和合法攻击。

·鲁棒性攻击:鲁棒性攻击以减弱或消除数字水印信息为目的。攻击手段主要包括压缩、剪切、模糊化、旋转、再抽样、平滑、滤波等。

·表示攻击:表示攻击不一定删除水印,而是通过对数字作品的内容进行处理,使检测器找不到有效的水印信息。典型的攻击手段有几何变形攻击和马赛克攻击。

·解释攻击:解释攻击通过向数字作品中嵌入伪造的水印,由于无法检测水印嵌入的先后顺序,造成原来的水印不能用于判断数字作品的所有权。

·合法攻击：合法攻击主要利用法律问题的漏洞。

S. Voloshynovskiy[6]从水印系统性能评估角度，依照Stirmark的评估模型，引入了另一类攻击——预测攻击，它根据原始图像和差值图像的统计分布特性，从含水印图像中预测原始图像，从而达到删除、削弱或伪造数字水印的目的，并把它作为第二代水印系统评估方法。其实，预测攻击问题也就是如何从含水印图像中消除因嵌入水印引入的噪声。

3. 去噪攻击的数学描述和对策

去噪攻击不需要知道水印嵌入的规则，利用原始图像和水印信息统计特性的先验知识，从含水印图像中预测原始图像或水印。参照图1，当攻击者得到含水印图像 y 后，把嵌入的水印信息作为原始图像的噪声，不考虑嵌入过程中所采用的水印算法，含水印图像可表示为：

$$y=x+n \tag{2}$$

其中 n 是水印序列 w 采用扩频技术嵌入原始图像过程中引入的噪声，也就是含水印图像和原始图像的差值图像。在没有原始图像信息的情况下，可以通过原始图像的估计值 x^* 和含水印图像 y 的差来估计噪声 n 的值：

$$n^*=y-x^* \tag{3}$$

由于只有含水印图像 y 是已知的，因而从含水印图像 y 估计原始图像 x^* 的问题可以归结为图像去噪的问题，这样可以利用已经发展比较成熟的除噪理论。

3.1 基于极大似然(ML)估计的图像去噪方法

如果没有原始图像统计分布特性的先验知识，可以假定原始图像是均匀分布的，差值图像的统计分布为 $p_n(n)$，这样可以使用ML方法对原始图像进行估计：

$$x^*=\arg\max_{\tilde{x}\in R^N}\{\ln p_n(y/\tilde{x})\} \tag{4}$$

根据文[7,8]，在图像恢复和去噪过程中，差值图像的统计特性可以用非平稳高斯模型或平稳广义高斯模型进行描述。在此，只考虑平稳广义高斯模型的两种特殊情况：平稳高斯分布和拉普拉斯分布。

当差值图像服从平稳高斯分布时，原始图像的最大似然估计 x^* 可以由 y 的局部平均值给出：

$$x^*=Localmean(y) \tag{5}$$

如果差值图像服从拉普拉斯分布，原始图像的最大似然估计 x^* 可以由 y 的局部中值给出：

$$x^*=Localmedian(y) \tag{6}$$

数字图像处理中，局部平均值是一种平滑线性滤波器，通过与模板卷积来实现，卷积对应频域中的乘积，局部平均值其实是频域中的低通滤波器。局部中值是一种中值非线性滤波器，与平滑线性滤波器相比，它能在衰减随机噪声的同时，保留图像的边界，从频域角度来看它和低通滤波相似，如图2所示。

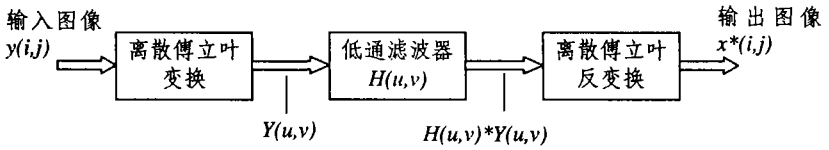


图2 低通滤波器

根据傅立叶变换的性质，一幅图像中，边缘、纹理和尖锐的跳跃(例如噪声)对傅立叶变换的高频分量有很大贡献，低通滤波器对一定范围内的高频分量进行衰减能够达到图像平滑化和去噪的效果。含水印图像 y 是由原始图像 x 和差值图像 n 叠加得到的，差值图像 n 是含水印图像 y 的噪声。从频域角度来说，为了不让差值图像 n 被低通滤波器滤除，那么差值图像 n 的能量谱必须集中在低通滤波器低频带通区域，滤波后就能保持差值图像的谱能量，这样算法抗低通滤波的性能才好。因而我们引入低通滤波后差值图像损失的谱能量与低通滤波前差值图像的谱能量的比值来度量算法抗低通滤波的性能：

$$\text{谱能量衰减经值}=\frac{\text{滤波前谱能量}-\text{滤波后谱能量}}{\text{滤波前谱能量}} \tag{7}$$

如果谱能量衰减比值越大，算法抗低通滤波的性能越差；如果谱能量衰减比值越小，算法抗低通滤波的性能越好。

3.2 基于最大后验概率(MAP)估计的图像去噪方法

如果已知原始图像 x 和差值图像 n 统计分布特性的先验知识，那么可以用MAP方法对原始图像进行估计。假设原始图像的统计分布为 $p_x(x)$，(2)式对原始图像的MAP估计为：

$$x^*=\arg\max_{\tilde{x}\in R^N}\{\ln p_n(y/\tilde{x})+\ln p_x(\tilde{x})\} \tag{8}$$

如果原始图像服从于局部均值为 $\bar{x}_j$，局部方差为 $\sigma_{x_j}^2$ 的非平稳高斯分布，即 $x\sim N(\bar{x}_j,\sigma_{x_j}^2)$，差值图像 n 服从于均值为0，方差为 σ_n^2 的高斯分布，那么(8)式的解为自适应维纳滤波器：

$$x^*=\bar{y}+\frac{\sigma_x^2}{\sigma_x^2+\sigma_n^2}(y-\bar{y}) \tag{9}$$

如果原始图像 x 服从于平稳的拉普拉斯分布，均值为 $\bar{x}$，方差为 σ_x^2，差值图像 $n\sim(0,\sigma_n^2)$，(8)式的解为：

$$x^*=\bar{y}+\max(0,|y-\bar{y}|-\sqrt{2}\,\sigma_n^2/\sigma_x)sign(y-\bar{y}) \tag{10}$$

我们知道，对于图像的平滑区域，图像的局部方差趋向于0，对于图像的纹理或边缘区域，图像局部方差值比较大。上述两种滤波器将图像 y 分解为低频部分 $\bar{y}$ 和高频部分 $(y-\bar{y})$，当 $(y-\bar{y})$ 的值较小时，表示平滑区域；$(y-\bar{y})$ 值较大时表示纹理或边缘。通过分析两个滤波器的工作原理，不难看出它们去噪主要是通过抑制平滑区域的噪声来实现的。因此，如果差值图像的区域分布特性和原始图像区域的分布特性很相似，滤波对差值图像能量的衰减就会很小，那么上述三种滤波器就很难将差值图像的信息滤除。由于局部方差反映了图像的区域分布特性，因而可以通过分析差值图像的局部方差和原始图像的局部方差分布的相似性来检验水印算法抗MAP图像去噪方法的性能。为此我们采用均方误差的方法衡量差值图像和原始图像局部方差分布的相似性：

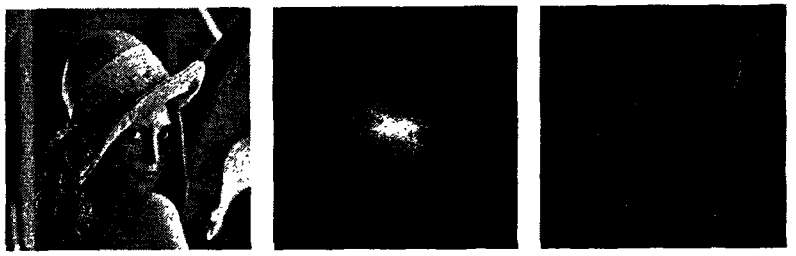
$$MMSE=\frac{1}{M*N}\sum_{i=1}^{M}\sum_{j=1}^{N}[dev_n(i,j)-dev_x(i,j)]^2 \tag{11}$$

4. 试验结果

为了验证上述理论分析结果，我们采用cox[5]、kim[3,4]、wang[2]和xia[1]的水印嵌入方法进行测试。其中Cox方法在DCT变换域嵌入水印，xia、wang和kim是小波变换域水印嵌入算法。采用的图像是512 * 512的lena标准图像，分别采用上述方法嵌入水印。

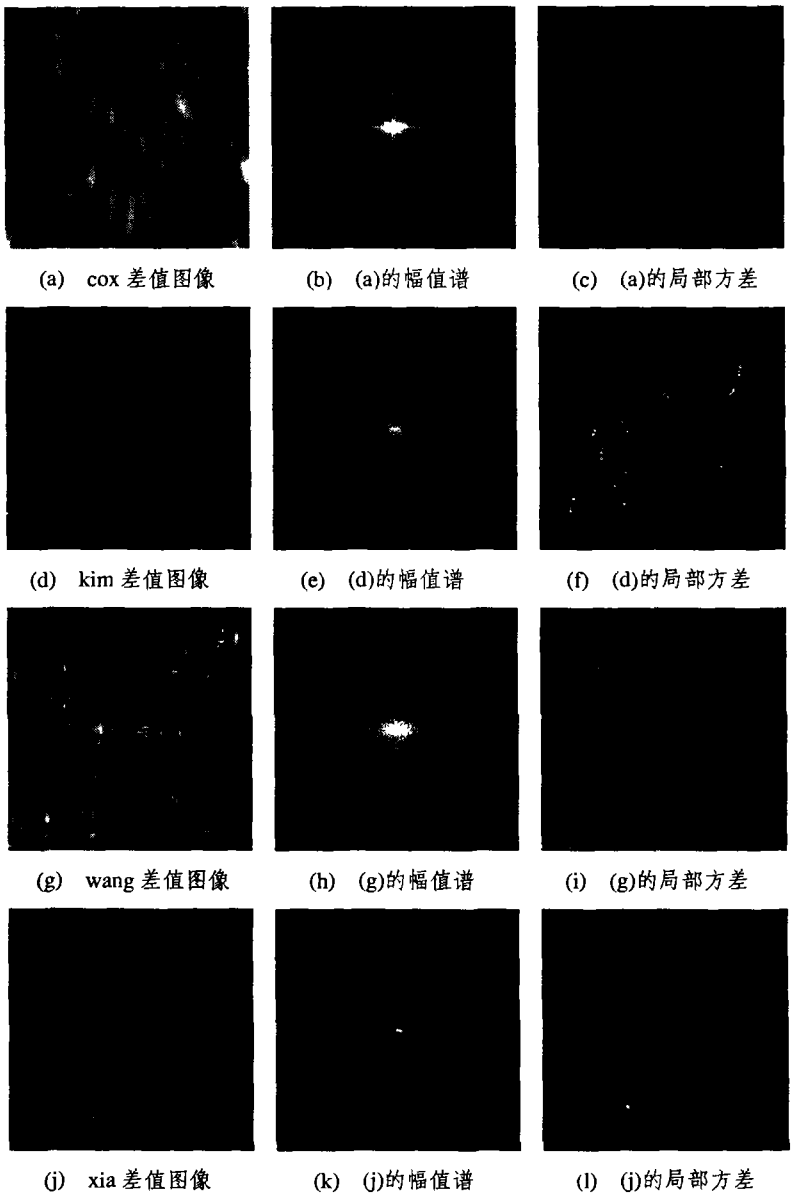
原始图像及其幅值谱、局部方差分布如图3所示。

嵌入水印后的差值图像及其幅值谱、局部方差分布如图4所示。



(a) lena 图像　(b) lena 图像幅值谱　(c) lena 图像局部方差

图3　原始图像及其幅值谱、局部方差分布



(a)　cox 差值图像　(b)　(a)的幅值谱　(c)　(a)的局部方差

(d)　kim 差值图像　(e)　(d)的幅值谱　(f)　(d)的局部方差

(g)　wang 差值图像　(h)　(g)的幅值谱　(i)　(g)的局部方差

(j)　xia 差值图像　(k)　(j)的幅值谱　(l)　(j)的局部方差

图4　差值图像及其幅值谱、局部方差分布

对含水印图像进行局部平均和中值滤波，其中局部平均采用16 * 16的邻域平均值，中值滤波采用5 * 5的模板，滤波后分别计算差值图像的谱能量衰减比值。四种水印算法检测输出结果见表1。

可以看出，不同低通滤波器对差值图像谱能量衰减比值能够很好地度量差值图像谱能量在低频带通区域集中的程度，并能很好地评估水印算法抗相应低通滤波器的性能。

为了测试算法抗 MAP 估计图像去噪的性能，我们仍然采用 cox、kim、xia 和 wang 的水印嵌入方法，计算差值图像局部方差分布和原始图像局部方差分布的均方误差，然后对含水印图像分别进行自适应维纳滤波。四种水印算法检测输出的相关值结果如表2所示。

试验结果验证了上述分析得到的结论：即差值图像的局部方差和原始图像的局部方差分布的相似性能检验水印算法

抗 MAP 图像去噪方法的性能。

表1 抗低通滤波试验结果

水印算法	未处理前输出相关值	谱能量衰减比值	局部平均后输出相关值	谱能量衰减比值	中值滤波后输出相关值
Cox	1.0	0.016	0.9590	0.0048	0.9963
Kim	0.9995	0.242	0.2147	0.337	0.2498
Wang	0.9999	0.0465	0.4424	0.0519	0.5748
Xia	1.0	0.3247	0.1339	0.5556	0.1668

表2 维纳滤波试验结果

水印算法	均方误差($\times 10^5$)	Wiener 滤波后输出相关值
Cox	0.4747	0.9989
Kim	1.5329	0.5062
Wang	1.5293	0.8458
Xia	1.451	0.8965

结论 本文从水印攻击的角度，分析了去噪攻击的原理，得出了如下结论：如果差值图像的谱能量在低通滤波器低频带通区域越集中，差值图像谱能量衰减比值越小，那么差值图像被低通滤波滤除的可能性越小，算法抗 ML 估计去噪攻击的性能也越好；如果差值图像和原始图像的局部方差分布越相似，那么算法抗 MAP 估计去噪攻击的鲁棒性越好。相反则水印算法抵御去噪攻击的鲁棒性越差。通过实验我们证明了结论的正确性，实际工作中，可以利用上述结论有效地增强水印算法抗去噪攻击的能力，从而设计出性能更好的水印算法。

参考文献

1 Xia X-G, Boncelet C G. Wavelet transform based watermark for digital images. Optics Express, 1998, 3(12): 497

2 Wang H-J, Su P-C, Kuo C-C J. Wavelet-based blind watermark retrieval technique. In: Proc. of the SPIE Phonics East Symposiumon Voice, Video and Data Communications, Boston, MA, USA, 1998, 3528: 440～451

3 Kim J R, Moon Y S. A robust wavelet-based digital watermark using level-adaptive thresholding. In: Proc. of the 6th IEEE Intl. Conf. on Image Processing, ICIP'99, page 202, Kobe, Japan, Oct. 1999

4 Kim Y-S, Kwon O-H, Park R-H. Wavelet based watermarking method for digital images using the human visual system. Electronic Letters, 1999, 35(6): 466～467

5 Cox I, Kilian J. Secure spread spectrum watermarking for multimedia. IEEE Trans. Image Processing, 1997, 6: 1673～1687

6 Voloshynoveskiy S, Pereira S. Attack modeling: towards a second generation watermarking benchmark. Signal Processing of the ACM, 2001, 81: 1177～1214

7 Voloshynoveskiy S, Deguillaume F. Content Adaptive watermarking based on a stochastic multiresolution image modeling, EUSIPCO 2000, Tampere, Finland, Sep. 2000

8 Voloshynoveskiy S, Herrigel A. A stochastic approach to content adaptive digital image watermarking. In: Third Intl. Workshop on Information Hiding, Dresden, Germany, Sep. 29～Oct. 1, 1999

9 刘振华，尹萍. 信息隐藏技术及其应用. 科学出版社

（上接第121页）

强，这就是我们在使用此类网络(GMNN)最关心的，前面的分析以及实验结果说明确实存在某种最优的量化。

表1 实验结果

网络结构		RMS（训练集）	RMS（测试集）
SN-TUPLE1	L=512 m=10 n=31 M=500 d=8	0.0186	0.0848
	L=512 m=50 n=15 M=128 d=2	0.0287	0.1052
	L=512 m=50 n=15 M=128 d=4	0.0158	0.0585
	L=128 m=10 n=15 M=128 d=6	0.0835	0.1338
SLLUP	L=128 m=10 n=15	0.0783	0.1169
	L=512 m=10 n=31	0.0113	0.1124
ASDM2	M=500 D=5	0.0393	0.0874
	M=500 D=8	0.0463	0.0834
	M=900 D=8	0.0426	0.0817

结论 我们避开 CMAC、SLLUP 的具体的量化方式，它们按各自的方式实现从原始数据到地址空间的映射，其后的学习就是一个线性优化问题了，而非线性性质就隐含在上述这种映射之中。这一特点可以延伸到一大类称为一般存储器神经网络(GMNN)框架中，并且证明了它通过误差迭代学习能收敛到平方误差最小。由于 GMNN 的性能主要取决于如何量化以及量化程度，而因此决定了中间内部表示(即地址指示向量)。分析表明，确实存在(可以找到)一种量化方式，对于任意一个给定的学习问题，可以最小的网络规模达到目标，并给出了一个一般性的描述。当然，尚待研究的问题很多，对那类基于存储器的网络究竟作何评价、与其它普通的网络的关系，以及量化方式蕴含的非线性性质如何通过抽象出来的中间内部表示来揭示(已经有人作了部分工作)，还有过拟合(过学习)的问题等等，这些问题都会影响到这类网络的广泛应用。

参考文献

1 Funahashi K. On the approximate realization of continuous mappings by neural networks. Neural Networks, 1989, 2: 183～192

2 White H. Learning in artificial neural networks. Neural Comput. 1989, 1(4): 425～464

3 Albus J S. A theory of cerebellar function. Mathematical Biosciences, 1971, 10: 25～61

4 Bledsoe W, Browning I. Pattern recognition and reading by machine. IRE Joint Computer Conference, 1959. 225～232

5 Aleksander I, Thomas W V, Bowden P A. WISARD-a radical step forward in image recognition, Sensor Review, July, 1984. 120～124

6 Kolcz A, Allinson N M. N-tuple Regression Network. Neural Networks, 1996, 9(5): 855～869

7 Kolcz A, Allinson N M. General memory neural network-extending the properties of basis networks to RAM-based architectures. In: Proc. 1995 IEEE Int. Conf. on Neural Networks (ICNN'95), 1995. 1638～1643

8 Kolcz A, Allinson N M. Basis function models of the CMAC network. Neural Networks, 1999, 12: 107～126

9 Kolcz A, Allinson N M. The general memory neural network and its relationship with basis function architectures. Neurocomputing, 1999, 29: 57～84

10 Rohwer R, Morciniec M. The Theoretical and Experimental Status of the n-tuple Classifier. Neural Networks, 1998, 11(1): 1～14

11 Tattersall G D, Foster S, Johnston R D. Single-layer lookup perceptrons. IEE PROCEEDINGS-F, 1991, 138(1): 46～54

12 Wong Y-F, Sideris A. Learning convergence in the Cerebellar model articulation controller. IEEE Trans. On Neural Networks, 1992, 3(1): 115～121

13 Lin C-S, Chiang C-T. Learning convergence of CMAC technique. IEEE Trans. On Neural Networks, 1997, 8(6): 1281～1292

14 Kanerva P. Sparse Distributed Memory. MIT Press, Cambridge, Massachusetts, 1988