计算机科学2002Vol. 29№. 12

易损数字水印技术:研究与应用

Fragile Digital Watermarking: Research and Application

陆唯杰　陈克非

(上海交通大学计算机科学与工程系　上海200030)

Abstract Recently, digital watermarking techniques for multimedia get fast growth. Digital watermark is some information that will bring no perceptual feeling when embedded in the multimedia data. There are mainly two kind of watermark: robust watermark for copyright protection and fragile watermark for data integrity and data authenticity. In this paper, we will mainly focus on fragile watermarking. For the purpose of multimedia integrity and authenticity protection, we classify current methods into three categories: digital signature techniques, fragile watermarking techniques based on Hash function and other fragile digital watermarking techniques. And we discuss each category in this paper, and especially discuss fragile watermarking in detail. Such a classification will help us understand the functions of digital signature and fragile watermarking in data integrity and authenticity protection. And we also discuss some attacks on fragile watermarking and its applications.

Keywords Fragile watermark, Digital watermark, Alteration detection

1. 引言

随着数字化和网络化的发展,出现了数字多媒体的版权和完整性保护等问题。为了解决这些问题,人们开始做各方面的研究,而数字水印技术是其中的一个主要研究方向,并且得到了越来越多的关注。当前研究的数字水印,按其特性主要可以分为两类:鲁棒数字水印和易损数字水印。鲁棒数字水印(Robust Watermarking),要求对多媒体数据进行操作时,不会或者很难去除或破坏嵌在数据中的数字水印,也就是要求保证水印的鲁棒性。主要用于数字产品的版权保护,比如在数字作品中表示著作权信息,如作者、作品序号等,它要求嵌入的水印能够经受各种常见的编辑处理和各种恶意攻击。为了达到鲁棒性,有些方法是在变换域上嵌入水印[1,2];有些技术基于扩散频谱方法,将一个窄带信号(水印)通过一个有噪宽带信道(多媒体内容)来传输[2,3];还有一些具有图像可适应性的水印方法,比如根据图像的局部特性改变水印嵌入参数,这样可以在保证水印引入最小扭曲的同时最大化水印的鲁棒性[4]。

易损数字水印(Fragile Watermarking),与鲁棒水印的要求恰恰相反,易损水印要求多媒体数据被改动时,内嵌的数字水印很容易被破坏,也就是要求对信号的改动很敏感。主要可以用于真实性、完整性保护,根据易损水印就可以判断数据是否被篡改过,有时候还要求可以根据易损水印检测出改动发生的位置和改动的程度。用于完整性真实性保护的易损水印如 Yeung-Mintzer 提出的一种易损水印,对图像的线性或非线性变化都会改变内嵌的水印[5]。但是这样一类水印方法对于一些"信息保持"(information preserving)[11]操作来说并不理想,"信息保持"是指那些保持原先的图像表示内容和含义的操作(如图像压缩),这类操作不属于恶意的篡改,但是同样会破坏水印。所以有些人提出了半易损水印的概念,它要求水印既具有一定的鲁棒性,可以承受一些操作(如压缩),而对其他的操作又具有易损性。Wolfgang 和 Delp 提出的基于 VW2D 的易损水印方案[6]就是一种半易损水印,其他如文[11]等。半易损水印的目的还是为了保护数据的完整性和真实性,还是一种易损水印,所以在本文中不将其另外作为一种水印类型讨论。

本文将对易损水印进行讨论,主要从数据完整性真实性的角度出发,分类讨论数字签名以及易损水印技术的不同实现方法和技术。也讨论了一些攻击方法,并在最后描述了易损水印的一些应用以及其应用框架。

2. 易损数字水印技术

对于多媒体数据完整性、真实性的保护主要有两种方法:一是基于数字签名技术[7],提供完整性和真实性保护;二是利用数字水印技术[8],提供一种有利于应用的方法,比如它可以方便图像的管理,也可以定位图像的修改位置、修改程度等。具体来说数据完整性、真实性有三种保护方法:1)直接基于数字签名的方法,由于它不允许数据任何的一点改动,不适合多媒体应用,所以有人提出扩展用于签名的 Hash 函数来获得一定的鲁棒性,还有基于图像内容特征的签名及连续性真实(continuous authenticity)[10]的数字签名方法来得到相应的鲁棒性。2)将数字签名和水印技术结合的方法,就是先利用 Hash 函数对数据进行签名,然后利用水印技术将签名嵌入多媒体文件中。如 P. Wong 对图像,图像的大小以及水印嵌入时的密钥求 Hash 值,然后将得到的 Hash 值嵌入原始文件的 LSB(least-significant bit)层[9]。3)不依靠数字签名,利用其他方法保证完整性真实性的易损水印技术。如 Yeung-Mintzer 的基于空间域的易损水印方法[5];还有 Wu-Liu,为了支持 JPEG 压缩,在 DCT 域上的水印方法[12]。

2.1 数字签名技术

传统的基于密码学的数字签名技术也可以保护数据的完整性真实性,对整个文件内容通过 Hash 函数(如 MD5等)作个摘要,如果安全需要,可以再对这个摘要加密。验证时,对测试文件作同样的摘要,如果得到的摘要与解密得到的摘要相

陆唯杰　硕士研究生,研究方向为信息安全。陈克非　教授,博士生导师,研究方向为数据加密、网络安全、安全电子商务。

同,说明文件内容没有改变。这种方法用于多媒体数据时会带来一些不利的方面,首先,要求文件内容不允许有1比特的错误,而这个要求对于多媒体来说是不合适的,因为多媒体文件可能有压缩、去噪等操作,这些操作未必是恶意篡改,但是得到的文件却不能通过验证;其次,产生的摘要需要另外存放,虽然可以放在文件的头尾,但是同样会给多媒体文件的管理和应用带来相当大的麻烦。Friedman 提出的可信任的数码照相机[7],是基于数字签名技术的,对于密钥、摘要的管理还是需要很多的问题要解决。

Schneider 提出了用于图像验证的基于图像内容的鲁棒的数字签名技术[10],这样一种数字签名技术可以允许指定的一些图像操作(如有损压缩)但是不允许其他操作。他提出了连续真实(continuous authenticity),即如果两幅图像的每个像素都相同,则真实度为1.0;如果两幅图像什么都不相同,则真实度为0.0;其他的都是部分真实。真实度可以定义在图像特征的基础上,设 I', I"为两幅图像,基于特征的可信度 Af 可以有如下定义,Af = 1-‖feature (I')-feature (I")‖normalized。图1的真实度-修改度的曲线可以看出对不同的操作,有不同的曲线。而实现这个方法的目标就是找到相应的特征,使得对于某些操作(如有损压缩),这个曲线是平滑的;而对于不希望的操作(如剪切),曲线是陡的。

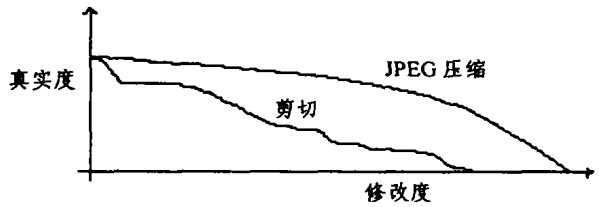

图1

Schneider 签名方法的产生和验证如图2,首先根据图像内容提取特征得到 Co,然后作 Hash 操作以减少数据量,得到 Ho,最后根据私钥 Kpr 对 Ho 加密,得到签名 S。

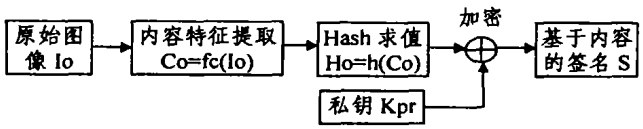

图2 生成签名

签名的验证过程如图3,对于测试图像提取特征 Ct,然后求 Hash 值,得到 Ht,根据公钥解密另外收到的签名 S,得到 Ho,然后比较 Ht 和 Ho。

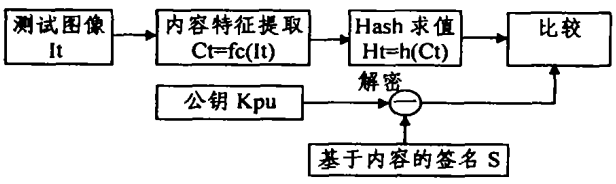

图3 验证签名

2.2 基于 Hash 函数的易损水印

这种类型的水印很大程度上与前一部分数字签名技术一样,只是最后在保存签名时利用了水印技术嵌入在数据中。就象 Wolfgang 提到的那样,简单地对整个图像做个 Hash 操作就可以作为一个易损水印了[6]。在很多时候要求易损水印在没有原始数据的情况下,具有可以检测修改发生位置的能力,这也是一般的数字签名所没有的特性。在文[6]中,Wolfgang 基于图像提出了两种方法来达到相同的目的:

1. 行-列 Hash 函数技术 对一幅图像的行和列用 Hash 函数分别作摘要,然后将这些摘要保存下来;检测时,同样对测试图像的行和列作 Hash 函数,那么修改发生位置就可以根据行列 Hash 值不同来做出判断。虽然这种方法对于单个像素的修改都可以定位,但有时会有一些局限性,会把没有被改动的部分也认为是修改过的。如图4,图中 A,B 是两个被修改的部分,用这种方法会导致两个阴影部分也被认为是修改过的。这样的局限性限制了 Hash 函数技术在图像中的应用。

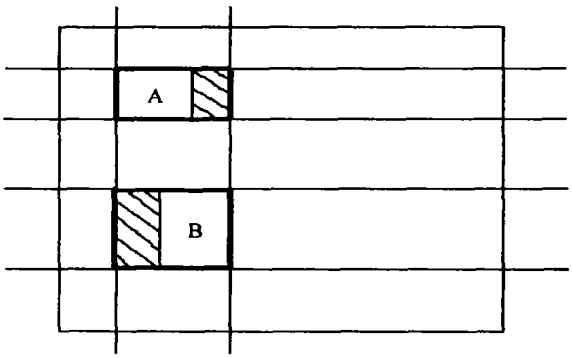

图4

2. 基于块的 Hash 函数技术 BBHF 这种方法是对图像分块,然后分别对不同的块求 Hash 函数值,并将这些结果保存,用来与测试图像得到的 Hash 函数值作比较,如果有不同,则就可以知道是图像中的那一块。

2.3 易损数字水印技术

按照水印的隐藏位置可以分为空间域数字水印和变化域数字水印,前者是直接在宿主数据信号空间上嵌入水印信息,变换域数字水印是在信号的 DCT 变换域、FFT 变换域、小波变换域等上隐藏水印。也有一些结合空间域和变换域的数字水印方法,如文[13,14]。

2.3.1 空间域数字水印 早期的易损水印都将水印直接嵌在数据的空间域上,如文[8,18],都是将水印嵌在图像的最低有效位 LSB(least significant bit),其修改对于图像效果的影响是最小的。但是这样的方法很容易被攻击,攻击者很容易修改一个图像而保持图像的 LSB 平面不变。这种方法还有一个缺点就是不能做任何的压缩操作,一做压缩,图像的 LSB 平面就会被破坏,使得图像的真实度变为不可信,而在实际应用中,为了传输或者存储的方便,压缩操作应被允许。

Yeung-Mintzer[5]和 Wong[9]提出了如下的水印方案,有一个随机密钥产生一个映射函数 f, $f:\{0,1,\cdots,255\}\rightarrow\{0,1\}$,就是将像素值0到255映射到0或1。如果是彩色图像则定义三个这样的函数:fr,fg,fb。这里要嵌入水印是一个二值标示图像,可以表示为 L(i,j)={0,1}。对于灰度图像,像素(i,j)的灰度值为 g(i,j),尽量小地修改这一灰度值使得对每个像素满足下式:L(i,j)=f(g(i,j))。

对于彩色图像,则尽量小地修改三个信道的颜色值,使得对每个像素满足:

L(i,j)=fr(R(I,j)) XOR fg(G(I,j)) XOR fb(B(I,j))。

其中 R,G,B 分别为图像的三个颜色分量。验证时,只要将函

数 f 作用于图像上就可以得到相应的标示图像：L(i,j)=f(g(i,j))。这种方法有几个好处，首先就是表示图像本身具有有用的视觉信息，这些信息可以是关于图像本身的或者其他信息；其次，从测试文件中得到的标示图像与原始的标示图像就可以看出是否有过修改，并且判断出修改的位置；再次，这种水印的嵌入，往往不仅仅是在 LSB 层，通常要深一点，这样增加了安全性；最后这种算法快速简单，比较容易用硬件实现。

在文[25,26]中，针对上述 Yeung-Mintzer 的水印方法提出了一些攻击方法，具体的攻击方法见第3节。Fridrich[24]在这些基础上提出了一种新的易损水印方法，这个方法主要针对文[25,26]的攻击。

Wolfgang 提出了用于验证图像的 VW2D(Variable-Watermark Two-Dimensional)算法[10]，这里的水印是一个随机数，水印的加入是基于块的。水印的嵌入过程，Y(b)=X(b)+W(b)，这里的 X(b)是原始图像块，W(b)是水印块，Y(b)是嵌入水印后的图像块，对每个图像块作如上操作。验证图像 Z 看水印是否位于其中，e=Y(b)*W(b)−Z(b)*W(b)，如果 e<T，Z(b)被认为是真的，这里的 T 是用户定义的下限。从验证过程可以看出，验证并不需要原始图像 X。在文[20]中，Wolfgang 扩展了上述关于 T 的测试，认为对一个图像块可以根据测试结果归为：1)没有修改；2)轻微修改；3)肯定修改了但还是来源于加入水印的图像；4)完全改动了或者不是来源于加入水印的图像。并且他也列出了一些结果表明 VW2D 可以承受一些不同的攻击。在文[6]中，Wolfgang 对如何将 VW2D 用作半易损水印作了说明，并且通过与基于块的 Hash 函数 BBHF 的实验比较，说明 VW2D 比 BBHF 对修改有更好的鲁棒性。这种方法比直接嵌入空间域有更强的鲁棒性，被称为半易损水印。

2.3.2 *变换域数字水印* 主要的变换域有离散余弦变换(DCT)、离散小波变换(DWT)等。基于变换域上的水印具有鲁棒性，变换域的方法对于易损水印也有很多好处，最重要的就是这样的水印可以经受有损压缩。

Wu-Liu[12]提出了一个基于 JPEG 压缩图像的易损水印方案，跟 Yeung-Mintzer[5]的映射函数一样，需要一个随机生成的查找表 LUT(look-up table)，不同的是这里是将 JPEG 的系数映射到{0,1}，而 Yeung-Mintzer 是将像素值映射到{0,1}。假设 v 为原始的系数，v' 为嵌入水印后的系数，b 为要嵌入的水印比特，LUT(.)为查找映射表函数。嵌入水印时，如果 LUT(v)=b，v'=v，否则 v'=v"，v"为偏离 v 最小且满足 LUT(v")=b 的系数值。提取水印时，b'=LUT(v')，b' 为提取的水印比特。为了保证嵌入水印后图像不能有视觉上的改变，系数的变化必须在一定范围内，所以 LUT 中连续的"1"和"0"的长度必须受到限制。

Lin[11]提出的半易损水印，它的嵌入是在 DCT 域完成的，而它的检测是基于空间域的。Lin 指出这种水印当中度压缩时，能够判断修改区域的精度是75%，当轻度压缩时，精度可以达到近90%。

Xie-Arce[21]的易损水印方法是基于离散小波变换的，首先利用 SPIHT 算法[23]将图像变换到压缩形式，然后有选择性地将水印比特图像处理的方式嵌入图像中。

3. 易损水印的攻击方法

易损水印是用于数据完整性和真实性保护的，那些想伪造或篡改数据的人总是想方设法对易损水印进行攻击。

最简单的是一些基本攻击，比如一些图像处理操作，做这些操作的人往往不知道水印的存在，这类攻击应该很容易被各种水印方法检测出。

还有就是修改图像本身，而不改动水印。就象前面提到的，水印嵌入在图像的 LSB 层，攻击者完全可以改动图像甚至替换一个图像而保持相同的 LSB 层。而 Yeung-Mintzer[5]的方法嵌入水印不仅仅是在 LSB 层，所以对于攻击者来说更困难作这样的攻击。

攻击也可能是将某一数据的已知水印用于其他的数据。这样的方法同样可以用于本身，首先将水印提取出来，然后改动数据，最后再把水印插入，这样篡改后的数据还是具有合法水印，当然这里必须假定攻击者知道水印的嵌入和提取方法。如果嵌入的水印还跟宿主数据内容有关的话，则会大大加强水印抵御这种攻击的能力。

还有些攻击从推断水印嵌入的密钥入手或者直接攻击水印的嵌入和检测机制，如文[24～26]提到的两类对 Yeung-Mintzer[5]的攻击：

1)如果同样的标示图像和随机映射函数 f 被反复用于几幅图，那么很容易就可以比较精确地推出标示图像和 f，攻击者就可以将标示图像利用 f 作用于其他的不真实图像来达到假冒真实图像的目的。

2)如果攻击者有好几幅加入相同水印的图像，而水印值和图像的局部特性相关(比如只取决于一个像素的颜色值)，那么攻击者就可以将不同图像的不同部分组成一幅图像。这种攻击的关键在于，不同图像的相同部分具有相同的水印，所以只要攻击者有一定数量的图像，通过组合这些图像的不同部分就可以构造出一幅具有一样水印的图像来。

在实际情况下，不可能设计出可以抵御所有攻击的算法，但是为了设计出更好的水印算法，必须了解现有的攻击方法。

4. 易损数字水印的应用

易损数字水印要求宿主数据经过线性或非线性变换而改动时，水印很容易被改动和破坏。易损水印对于信号改动的敏感性使得它不适合用于多媒体数据的版权保护，而适用于多媒体数据的验证。

4.1 易损水印的应用框架

一个易损水印系统的框架分为嵌入部分和检测部分，基本上与大多数的水印系统类似。水印的嵌入框架如图5。

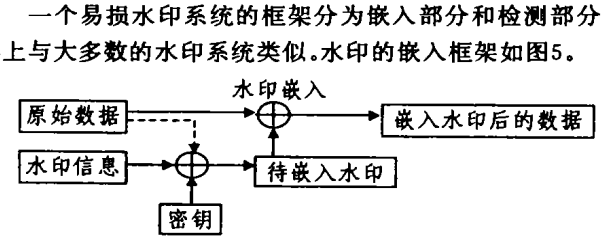

图5

密钥是用来对水印信息加密或者生成水印的关键部分，密钥要求保密管理，虚线表示水印的生成可能也取决于原始数据。水印的检测框架如图6。

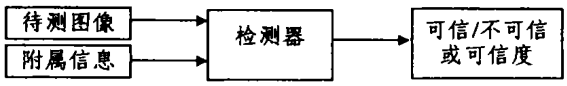

图6

这里的附属信息是相当重要的，包括水印嵌入时的密钥

等，必须对外界保密。最后检测器的输出结果为二值输出：图像可信或者不可信，也有可能是一个连续性的表示：图像的可信度。在这里我们必须考虑如何使得附属信息保密，而大部分实现的系统都认为这个问题可以以密码学的方法解决，或者认为存在完全可信的信道来传输这些信息；而在实际应用中，要真正解决这个问题，并不是很容易就可以实现的，这给易损水印的实际应用增加了不少障碍。

4.2 易损水印的基本应用

数据验证系统在法律、商业、国防、新闻等很多方面有很重要的应用。因为数字化数据很容易被修改，所以在很多情况下都会对一些数据的可信度表示怀疑，这时候如果能够判断数据是否被修改过是很有用的。一些例子如在数据库中的图像加入水印用于检测图像是否被篡改[15,16]，在贸易中购买者确信自己购买的图像收到时是真实的[17]，还有很多时候需要对视频或者音频的可信度提出怀疑，在这些方面易损水印都会有很大的应用。

虽然目前已经有了实际可用的数字水印系统，但是还是有很多研究机构和公司投入很大的人力物力在研究数字水印，这方面的研究还在继续。一方面在于这个技术本身缺少很强的理论作为基础；另一方面攻击者总是可以从不同的角度找到突破口，所以水印的设计者和攻击者之间的较量没完没了。数字水印是否可以作为法庭上的证据，这个问题还没有定论并且也很难下结论。就像前面提到过的那样，不可能有万能的数字水印方案。因此，从现阶段看，数字水印的设计应该针对每个特定的应用。找到一个需要数字水印技术的应用，分析可能要防范的攻击，根据攻击的破坏程度以及应用的需要进行取舍，最后设计出需要的水印方法。就像有些出版公司采用一些水印技术来保护版权，实际上他们并不是说要将这个作为法庭上的证据，而是作为一个威慑，比如说：如果你在网上随意发布他们出版的多媒体数据，他们可以知道。也可以说将水印技术用作一个发现非法行为的手段，因为发现之后可以通过其他途径和方法找到相应的证据。因此，可以相信作为数字水印技术中的一种，易损数字水印针对不同的特定应用环境会有相当广阔的应用前景。

结论 易损数字水印是将信号（水印）嵌在多媒体数据中，对数据的改动很敏感，也就是数据的改动很容易将水印破坏，这种水印被用来保护数据的完整性和真实性。本文主要讨论了现有的一些数据完整性真实性保护的方法，为了更好地说明数字签名、易损水印以及易损水印的鲁棒性和易损性与数据完整性真实性之间的关系，对它们做了分类讨论；也讨论了易损水印的应用和攻击方法。易损数字水印可以在很多重要的应用中得到运用。

参 考 文 献

1 Barni M, Bartolini F, Cappellini V, Piva A. A DCT-domain system for robust image watermarking. Signal Processing (Special Issue on Watermarking), 1998, 66(3): 357～372

2 Cox I, Kilian J, Leighton T, Shamoon T. Secure Spread Spectrum Watermarking for images, audio and video. ICIP, 1996

3 Tirkel A, Hall T. Advanced spread spectrum watermarking. In: Proc. of the ACM Multimedia and Security Workshop, Orlando, Florida, Oct. 1999. 37～42

4 Podilchuk C I, Zeng W. Image adaptive watermarking using visual models. IEEE Journal on Selected Areas in Communications (JSAC), 1998, 16(4): 525～539

5 Yeung M M, Mintzer F. An invisible watermarking technique for image verification. ICIP, 1997

6 Wolfgang R B, Delp E. Fragile watermarking using the VW2D watermark. In: Proc. of SPIE, Security and Watermarking of Multimedia Contents, San Jose, California, 1999. 25～27

7 Friedman G. The Trustworthy Digital Camera: Restoring Credibility to The Photographic Image. IEEE Transactions on Consumer Electronics, 1993, 39(4): 905～910

8 Walton, Steve. Image Authentication for a slippery new age. Dr. Dobb's Journal, April 1995. 18～26

9 Wong P. A watermark for image integrity and ownership verification. Final Program and Proceedings of the IS&T PICS 99, 1999. 374～379

10 Schneider M, Chang S. A Robust Content-based Digital Signature for Image Authentication. ICIP, 1996

11 Lin E T, Podilchuk C I, Delp E J. Detection of image alterations using semi-fragile watermarks. In: Proc. of the SPIE Intl. Conf. on Security and Watermarking of Multimedia Contents II, Vol. 3971, January 23 - 28, 2000, San Jose, CA

12 Wu M, Liu B. Watermarking for Image Authentication. ICIP, 1998.

13 Kutter M, Bhattacharjee S K, Ebrahimi T. Towards second generation watermarking schemes. ICIP, 1999, 1: 320～323

14 华先胜，封举富，石青云. 基于BCH编码的局部化标注水印算法. 模式识别与人工智能，2001

15 Mintzer F, Braudaway G, Yeung M. Effective and ineffective digital watermarks. In: Proc. of the IEEE Intl. Conf. on Image Processing, Santa Barbara, California, Oct. 1997. 9～12

16 Mintzer F, Braudaway G, Bell A. Opportunities for watermarking standards. Communications of the ACM, 1998, 41(7): 57～64

17 Wong P W. A public key watermark for image verification and authentication. In: Proc. of the IEEE Intl. Conf. on Image Processing, Chicago, Illinois, Oct. 1998. 455～459

18 van Schyndel R, Tirkel A, Osborne C. A digital watermark. In: Proc. of the IEEE Intl. Conf. on Image Processing, Chicago, Illinois, Oct. 1998, 2: 404～408

19 Wolfgang R B, Delp E J. A watermark for digital images. In: Proc. of the 1996 Intl. Conf. on Image Processing, Lausanne, Switzerland, 1996. 219～222

20 Wolfgang R B, Delp E J. Techniques for watermarking digital imagery: further studies. In: Proc. of the Intl. Conf. on Imaging Science, Systems, and Technology, Las Vegas, Nevada, USA, 1997. 279～287

21 Xie L, Arce G. Joint wavelet compression and authentication watermarking. In: Proc. of the IEEE Intl. Conf. on Image Processing, vol. 2, Chicago, Illinois, Oct. 1998. 427～431

22 Kunder D, Hatzinakos D. Digital watermarking for telltale tamper proofing and authentication. Proceedings of the IEEE, 1999, 87(7)

23 Said A, Pearlman W A. A new fast and efficient image codec based on set partitioning in hierarchical trees. IEEE Trans. on Circuits and Systems for Video Technology, Mar 1996.

24 Fridirich J, Goljan M, Baldoza A C. New fragile authentication watermark for images. ICIP, 2000

25 Memon N, Shende S, Wong P W. On the security of the Yeung-Mintzer Authentication Watermark

26 Fridrich J, Goljan M, Memon N. Further Attacks on Yeung-Mintzer Fragile Watermarking Scheme. SPIE Intl. Conf. on Security and Watermarking of Multimedia Contents, II, Vol. 3971, San Jose, CA, Jan. 2000