Pitfalls in Public Key Watermarking

PAULO S.L.M. BARRETO, HAE YONG KIM

Universidade de São Paulo, Departamento de Engenharia Eletrônica
Av. Prof. Luciano Gualberto, tr. 3, 158; CEP 05508-900, São Paulo (SP), Brazil
{paulob,hae}@lps.usp.br

Abstract. The cryptographic concept of digital signatures is well suited to image watermarking for authentication and integrity verification. However, the peculiar nature of watermarks demands special care in the application of digital signatures for this purpose. Here we cryptanalyze a recently proposed watermarking scheme and show how it can be strengthened to thwart our attacks, exploring the use of hashing function contexts and signature algorithms based on elliptic curve cryptography.

1 Introduction

The concept of *digital watermarks* has been defined to face (and counteract) enormous challenges such as ownership, integrity and authentication of digital images posed by the spectacular growth of networked multimedia systems in recent years. A digital watermark is a visually imperceptible, information-carrying signal embedded in a digital image. Such a signal should be able, in many circumstances, to be *publicly* detected/verified. The need for this property is obvious: claims of image ownership could accept or reject, and issues of image integrity or authenticity could be settled, without the image owner or originator having to unnecessarily disclose any private information.

A natural way to achieve this is by means of public key cryptographic algorithms. However, the peculiar nature of watermarks demands special design care in the application of digital signatures; all known cryptanalytical techniques must be kept in mind to avoid subtle flaws that invalidate the scheme.

We will concentrate here on the aspects of integrity and authenticity rather than that of ownership resolution.

2 Watermarks and Digital Signatures

Surprisingly, most proposed watermarking schemes are *a priori* constructions based on symmetric cryptography principles rather than on asymmetric (*aka* public key) algorithms, and explicitly preclude public verifiability of watermarks. Wong's construction[6] is one of the few published techniques to address the problem from the point of view of cryptographically strong digital signatures.

The Wong scheme for watermark insertion in grayscale images can be summarized as follows. Let Z be an $M \times N$ image to be watermarked, and A a publicly available bilevel image of the same size to be used as visually meaningful watermark. Partition Z (resp. A) into blocks Z_t (resp. A_t) of 8×8 pixels (at most; border blocks may be shorter), $t = 0 \ldots r - 1$. Each block will be watermarked separately. Let Z_t^* be the block obtained from Z_t by clearing the least significant bit of all its pixels. Using a cryptographically secure hashing function H, compute the message digest $H_t \equiv H(M, N, Z_t^*)$. Truncate H_t to s bits, where s is the number of pixels in Z_t (generally 64, but possibly less for border blocks). Exclusive-or the truncated digest with A_t, getting the marked digest $\hat{H}_t$. Encrypt $\hat{H}_t$ with the image owner's *private* key, thus effectively generating a digital signature S_t. It is implicitly assumed that S_t is of the same size as $\hat{H}_t$, i.e. s bits. Wong suggests using the well-known RSA algorithm for signing. Insert S_t into the least significant bits of Z_t^*, obtaining the marked block X_t.

For color images, the above algorithm is applied independently to each of the color planes. The corresponding watermark verification algorithm is straightforward.

3 Security and Efficiency Considerations

We now analyze cryptanalytical weaknesses, storage requirements and processing times of Wong's method, and suggest means to make it robust and more efficient.

The color image construction lacks the ability of detecting *any* changes in the marked image since it can be subverted by reordering the color planes or extracting one of them into a grayscale image. Because the watermark on each plane is independent of the others, it will remain unchanged and hence verifiable. This problem is easily overcome: sign colored blocks instead by feeding three monochromatic pixels at a time to the hashing function, each from a different color plane in some predefined order. This has the advantage of being three times faster than signing the color planes separately, since the main processing cost of digital signatures is due to asymmetric cryptography operations rather than hashing.

Another attack can be mounted against the independence of signed blocks. Suppose a malicious attacker has a

collection of *legitimately* watermarked images, all of them of the same size $M \times N$ and containing the same visually meaningful watermark A. Then it is possible to select blocks from the authentic images and, keeping the block coordinates, build with them a new image (of the same size) whose watermark will be falsely verified as legitimate. Depending on the characteristics of A, it might even be possible to swap blocks within a single marked image while keeping the embedded watermark unchanged. Our solution to this problem is to include *context information* in the message digests, i.e. define $H_t \equiv H(X_{t-1}, t, M, N, Z_t^*)$, with $X_{-1} \equiv 0$.

A digital signature scheme can be subverted if the hashing function can be made to *collide*. In the case of Wong's method, this means finding two image blocks B and $\bar{B}$ such that $H(M, N, B^*) = H(M, N, \bar{B}^*)$.

Let f be a function assuming n distinct and uniformly distributed values. Probed at q random and uniformly distributed points of its domain, f is likely to repeat values when $q = O(\sqrt{n})$. This purely stochastic behavior is called the *birthday paradox* (cf. [5], chapter 7), and does not depend on working details of f.

For the hashing function used in Wong's scheme (MD5 truncated to 64 or less bits) collisions are expected to occur after only $O(2^{32})$ steps, making it feasible to attack the system with relatively modest computational resources. Furthermore, given the current status of attacks against MD5[1], it is wise to choose an altogether different function like SHA-1[4] for this purpose wherever possible.

The only way to thwart birthday attacks is to use larger digests, but this poses problems in locating changes and handling images that cannot be exactly partitioned in whole blocks. To address these problems for grayscale images and m-bit hashing functions, we suggest a block size of at least $2m$ pixels, and hashing larger blocks at the image borders while embedding therein the same amount of watermark data.

An RSA encrypted message is of the same size as the RSA modulus. Since the block size in Wong's original scheme is only 64 pixels, a fitting RSA modulus cannot be larger than 64 bits, and so small a modulus can be factored even through trial division. On the other hand, a typical secure modulus size is 1024–2048 bits, but a 2048-bit RSA signature will only fit in a block of roughly 45×46 pixels, considerably degrading the resolution of image tampering localization. Moreover, practical images are likely to have to be partitioned into hundreds or thousands of individually signed blocks, creating a considerable processing overhead as digital signatures become computationally more expensive for larger keys.

To address these issues, we suggest using signature algorithms based on the discrete logarithm (DL) problem (cf. [5]), as they occupy only about $2m$ bits when used with an m-bit hashing function without decreasing the level of security. In addition, we recommend the use of elliptic curve versions of DL signature schemes. Elliptic curve cryptosystems[2] are quite attractive due to their high security level as compared to key length, which results not only in much smaller keys but also in noticeably reduced processing times.

The Nyberg-Rueppel[3] digital signature algorithm offers *message recovery* when the number of bits of the underlying group order is larger than the digest size; thus, it is very well suited for watermarking. The only drawback is that the amount of visually meaningful data that can be embedded is upper bounded by half the size of the whole signature. Fortunately this is only a concern for grayscale images, as the ideas of section 3 provide more space to embed the watermarks in colored images.

4 New Proposal

We have implemented in C/C++ a variant (for color images) of Wong's algorithm incorporating all changes suggested by the above analysis. We use the Nyberg-Rueppel signature scheme with SHA-1 as hashing function and an elliptic curve over $GF(2^{170})$ containing a cyclic subgroup of prime order $p \approx 2^{161}$. Host images are partitioned in blocks of 16×8 pixels. In each block we embed 128 bits of visually meaningful data through exclusive-or with the block digests. Our experiences show that 256×256 color images are marked about 4 times faster with our elliptic curve scheme (about 10s on a Pentium-233) than with 1024-bit RSA and independent color planes, even though the resolution in our scheme is 4 times more accurate and lacks all weaknesses described.

References

[1] H. Dobbertin, "The Status of MD5 After a Recent Attack," *CryptoBytes* 2(2), 1996, pp. 1–6.

[2] A.J. Menezes, "Elliptic Curve Public Key Cryptosystems," Kluwer, 1993.

[3] K. Nyberg and R. Rueppel, "A new signature scheme based on the DSA giving message recovery," *First ACM Conf. Comp. Comm. Security* (1993), ACM Press, pp. 58–61.

[4] FIPS PUB 180-1, *Secure Hash Standard*, U.S. Dept. of Commerce/NIST, April 17, 1995.

[5] D.R. Stinson, "Cryptography: Theory and Practice," CRC Press, Inc., 1995.

[6] P.W. Wong, "A Public Key Watermark for Image Verification and Authentication," 1998 IEEE Int. Conf. Image Proc., vol. I, MA11.07.