Security on Fragile and Semi-fragile Watermarks Authentication

Shilpi Saha
Computer Science and Engineering Department
Heritage Institute of Technology
Kolkata, India

Debnath Bhattacharyya
Department of Multimedia
Hannam University
Daejeon, Korea

Samir Kumar Bandyopadhyay
Department of Computer Science and Engineering
University of Calcutta
Kolkata, India

ABSTRACT

Digital image manipulation software is now readily available on personal computers. It is therefore very simple to tamper with any image and make it available to others. Ensuring digital image security has therefore become a major issue. Watermarking has become a popular technique for copyright enforcement and image authentication. The aim of this paper is to present an overview of some possible attacks which may cause harm the present watermarking techniques. Here we have identified the attacks which frequently attacks some very well known fragile and semifragile watermarking techniques..

General Terms

Security, Watermarking and Authentication.

Keywords

Attack, image authentication, watermark, fragile, semi-fragile, security, private key, public key.

1. INTRODUCTION

In today's world, digital images and video are gradually replacing their classical analog counterparts. This is quite understandable because digital format is easy to edit, modify, and exploit. Digital images and videos can be readily shared via computer networks and conveniently processed for queries in databases. Also, digital storage does not age or degrade with usage. On the other hand, thanks to powerful editing programs, it is very easy even for an amateur to maliciously modify digital media and create "perfect" forgeries. It is usually much more complicated to tamper with analog tapes and images. Tools that help us establish the authenticity and integrity of digital media are thus essential and can prove vital whenever questions are raised about the origin of an image and its content.

In the past few years, many new techniques and concepts based on data hiding or steganography have been introduced as a means for tamper detection in digital images and for image authentication. Fragile watermarks are designed to detect every possible change in pixel values. In many schemes, it can be shown that without the secret key, the probability of a modification that will not be detected can be related to a cryptographic element present in the scheme, such as a hash function. Semi-fragile watermarks are moderately robust and thus provide a "softer" evaluation criterion (authentication with a "degree"). Some schemes have been specifically designed to be compatible with certain distortion, such as JPEG or wavelet compression. There is another special group of authentication techniques that can be termed "content authentication". In those schemes, robustly extracted image features are embedded in the image in a semi-robust manner to help identify gross changes in the image.

In the security domain, an integrity service is unambiguously defined as one, which ensures that the sent and received data are identical. This binary definition is also applicable to images. In real life situations, images can be transformed, their pixel values can be modified but not the actual meaning of the image. In order to provide an authentication service for still images, it is important to distinguish between malicious manipulations, which consist of changing the content of the original image and manipulations related to the use of images.

In this paper, we focus on different attacks on different watermarking techniques. So many watermarking schemes have been designed to provide authenticity to the digital images. But many of the proposed ideas are not totally secure because of the attack threats. We have specially focused on the identification of the attacks to which the proposed watermarking schemes are vulnerable..

2. GENERIC IMAGE AUTHENTICATION SYSTEM

To be an effective image authentication system, it must satisfy the following criteria [8]:

a. Sensitivity: The system must be sensitive to malicious manipulations.

b. Tolerance: The system must tolerate some loss of information and more generally non-malicious manipulations.

c. Localization of altered regions: The system should be able to locate precisely any malicious alteration made to the image and verify other areas as authentic.

d. Reconstruction of altered regions: The system may need the ability to restore, even partially, altered or destroyed regions in order to allow the user to know what the original content of the manipulated areas was.

In addition, some technical features must be also considered:

i. Storage: authentication data should be embedded in the image rather than in a separated file.

ii. Mode of extraction: depending on whether authentication data is dependent or not on the image, a full blind or semi-blind mode of extraction is required.

iii. Asymmetrical algorithm: contrary to classical security services, an authentication service requires an asymmetrical algorithm.

iv. Visibility: authentication data should be invisible under normal observation.

v. Robustness and security: it must not be possible for authentication data to be forged or manipulated.

vi. Protocols: it is obvious that any algorithm alone can not guarantee the security of the system. It is necessary to define a set of scenery and specifications describing the operations and rules of the system.

3. GENERAL FRAMEWORK FOR WATERMARKING

Watermarking is the process that embeds data called a watermark or digital signature or tag or label into a multimedia object such that watermark can be detected or extracted later to make an assertion about the object. The image may be an image or audio or video.

In general, watermarking scheme must consist following three parts [3]:

- The watermark
- The encoder
- The decoder and comparator

Each owner has a unique watermark or an owner can also put different watermarks in Different objects.

Encoding process: Let us denote an image by I, a signature by S and the watermarked image by I'. E is the encoder function, it takes an image I, a signature S and it generates a new image which is called watermarked image I', mathematically,

$$E(I,S) = I'$$

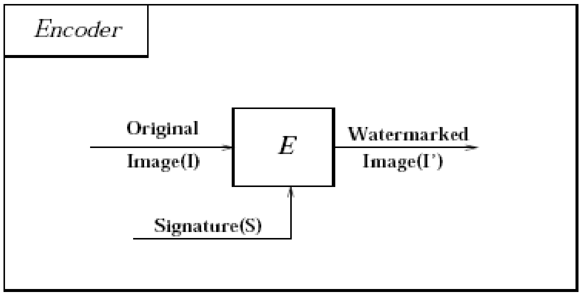


Figure 1. Encoder.

Decoding process: A decoder function D takes an image J, whose ownership is to be determined and recovers a signature S' from the image. In this process, an additional image I can also be attached which is often the original version of J. Mathematically,

$$D(J,I) = S'$$

The extracted signature S' will then be compared with the owner signature sequence by a comparator function $C\partial$ and a binary decision output is generated. It is 1 if there is a match and 0 otherwise.

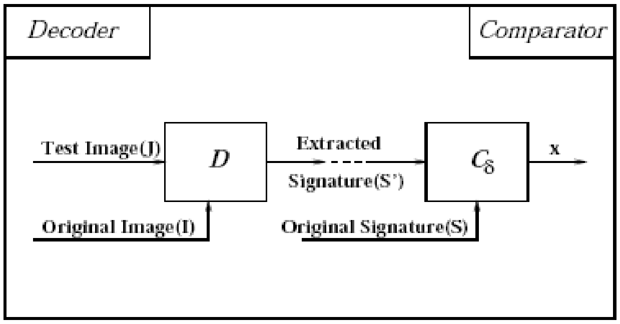


Figure 2. Decoder.

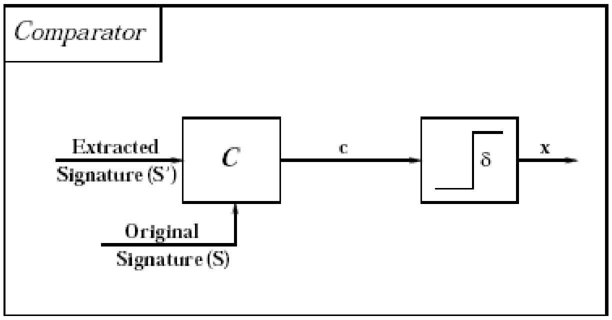


Figure 3. Comparator.

A watermark must be detectable and extractable. It should be noted that watermark extraction can prove ownership whereas watermark detection can only verify ownership.

4. TYPES OF WATERMARKING

Watermarks and watermarking techniques can be divided into various categories in various ways. The watermarks can be applied in spatial domain. An alternative to spatial domain is frequency domain watermarking. It has been pointed out that the frequency domain methods are more robust than spatial domain watermarking methods.

According to human perception, watermarking techniques can be divided into two categories: fragile watermarking and semi-fragile watermarking. We are discussing them:

a) Fragile Watermark: Basic idea behind this technique is to insert a specific watermark (generally independent of the image data) so that any attempt to alter the content of an image will also alter the watermark itself. Therefore, the authentication process consists of locating watermark distortions in order to locate the regions of the image that have been tampered with. The major drawback of this approach is that it is difficult to distinguish between malicious and non-malicious attack.

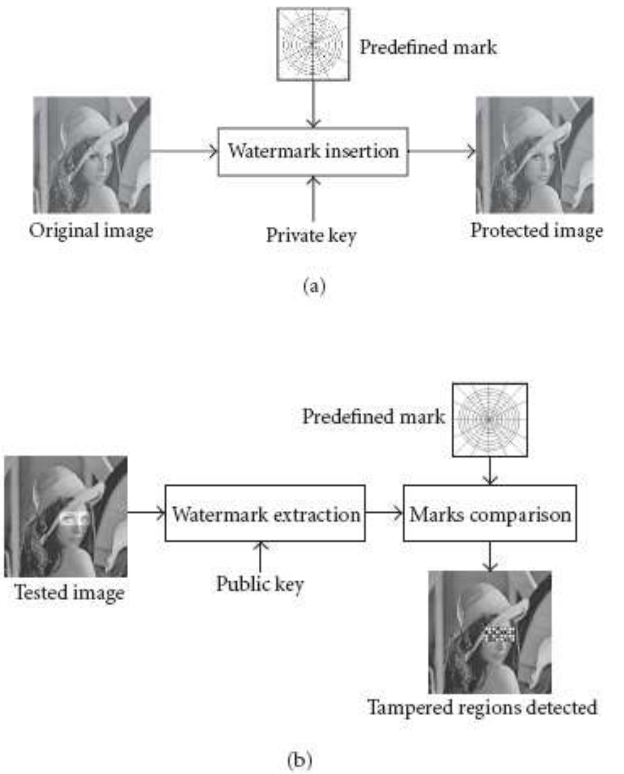
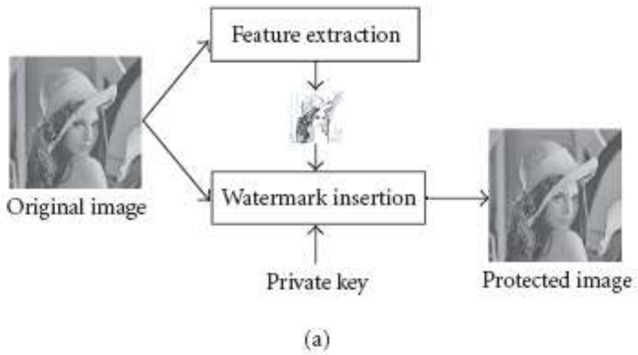


Figure 4. Fragile watermark scheme (a) image security (b) authenticity verification

b) Semi-fragile watermark: A semi-fragile watermark is another type of authentication watermark. Semifragile watermarks are more robust than fragile watermarks and less sensitive to classical user modifications. The aim of this method is to discriminate between malicious and non-malicious attack. The use of such method is justified by the fact that images are generally transmitted and stored in a compressed form.



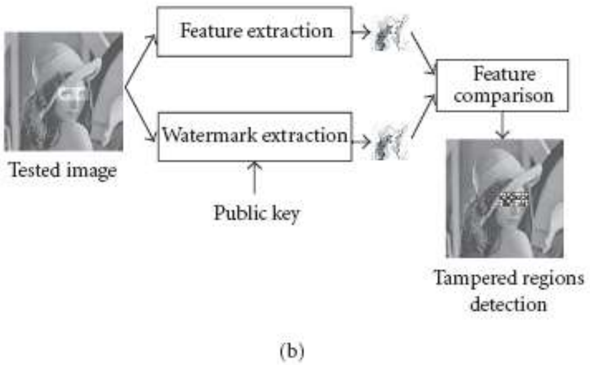


Figure 5. Semi-fragile watermark scheme (a) image security (b) authenticity verification

5. ATTACKS ON WATERMARKING

Actually, disturbing the watermark is quite easy because of its fragility. The goal of the attacker is the opposite when compared of a robust watermark. We further describe five attacks arranged in the increasing order of strength.

- Undetected Modifications: The attacker is trying to make a change to the authenticated image that will not be detected by the algorithm. He may even be satisfied with making changes that will not be detected with a "reasonable" probability or changes that will be misinterpreted by the detector, such as masquerading cutting and pasting as tampering at the border of the cropped area.
- Information Leakage: Another potential problem that many authentication watermarks have is information leakage. The attacker may be interested in obtaining some information about the secret authentication key, including the placement of MAC in the image pixels, detecting synchronization patterns, deriving portions of look-up tables, or obtaining some statistical evidence about the secret key or entities derived from it, such as a random walk through the image.
- Protocol Weakness: Even a scheme that does not have any information leakage and detects all modifications with very high probability may be vulnerable in certain situations. For example in the situation when an attacker has unlimited access to the verification device and is able to submit images for integrity verification.

The ability of an attacker to mount a successful attack depends on the specific application scenario in which the authentication scheme is used. In some applications, certain attacks may be irrelevant. Below, we divide the attacks based on the information and capabilities available to the attacker.

- Stego-Image Attack: The attacker has only one authenticated image and is interested in making changes that go undetected or recovering some secret information from the scheme.

- Multiple Stego-Image Attack: The attacker has multiple authenticated images and is interested in making undetected changes or recovering information from the scheme.
- Verification Device Attack: The attacker has access to the verification device, i.e., the attacker can verify the authenticity of any image. The strength of this attack depends on the output available to the attacker. The output could be a binary Yes/No for the whole image or it could be a bitmap with pixels/blocks indicated as authentic or tampered. Again, the attacker is interested in making undetected changes or recovering secret information from the scheme.
- Cover-Image Attack: The attacker has multiple pairs of original-authenticated images. This assumption is not that unreasonable if an attacker can somehow get access to the raw images before authentication has occurred or when plausible statistical hypothesis can be made about the original. Image semantic can also be used to obtain an estimate of the original image. Again, the attacker is interested in making undetected changes or recovering information from the scheme.
- Chosen Cover-Image Attack: The attacker has access to the authentication device and can submit his images for authentication.

6. PREVIOUS WORKS

In this section, we analyze the security of some current fragile watermarking and semi fragile watermarking schemes designed for authentication. The classification of attacks proposed in the previous section is used to evaluate the security and strength of the techniques and their usability.

Yeung-Mintzer scheme: One of the most popular (and most attacked) fragile watermarking schemes is the Yeung-Mintzer scheme [4]. The watermarking starts with a secret key that is used to generate a binary valued function f: $\{0, 1, \ldots, 255\} \rightarrow \{0,1\}$, that maps each grayscale level g_{ij} in the range from 0 to 255 to either 1 or 0. For color images, three such functions, f_R, f_G, f_B, one for each color channel, are generated. These binary functions are used to encode a binary logo L. The logo should be kept secret and can also be generated from the secret key or it can have graphical meaning. The gray scales g_{ij} are perturbed to satisfy the following expression for each pixel (i,j)

$$L_{ij} = f_g(g_{ij}) \qquad (1)$$

For an RGB image, all three color channels are perturbed to obtain

$$L_{ij} = f_R(R_{ij}) \oplus f_G(G_{ij}) \oplus f_B(B_{ij}) \qquad (2)$$

where $\oplus$ denotes the excluded OR and R_{ij}, G_{ij}, and B_{ij} are the values of the red, green, and blue channels, respectively. The pixels are updated sequentially in a row-by-row manner to enable error diffusion to better preserve the original colors. The image authenticity is easily verified by checking the relationship $L_{ij} = f_g(g_{ij})$ for each pixel (i,j).

There are some obvious advantages of this approach. First, the logo itself can carry some useful visual information about the image or its creator. It can also represent a particular authentication device or software. Second, by comparing the original logo with the recovered one, one can visually inspect the integrity of the image. Third, the authentication watermark is embedded not only in the LSBs of the image but somewhat deeper (± 5 gray scales). This makes it more secure and harder to detect. Fourth, the method is fast, simple, and amenable to fast and cheap hardware implementation. This makes it very appealing for still image authentication in digital cameras. It has also excellent localization accuracy because each pixel is individually watermarked.

There is a 50% chance that, in a non-watermarked image, for any given pixel (i,j) the expression (1) or (2) will be satisfied. Thus, without the knowledge of the binary functions and the logo, the probability that modifying n pixels in the watermarked image will not be detected will decrease exponentially with n. This degree of security may not be enough for some applications where it is important that all changes are detected with very high probability.

i) Wong scheme: One of the first fragile watermarking techniques proposed for detection of image tampering was based on inserting check-sums of gray levels determined from the seven most significant bits into the least significant bits (LSBs) of pseudo-randomly selected pixels. In this section, we describe the variation by Wong [5] because our new method is based on this technique.

Wong divides the image into non-overlapping blocks of $W \times H$ pixels. The watermarking is done for each block separately. Wong described two versions of this algorithm: private key and public key versions. In the private key version, the seven most significant bits of all pixels in the block are hashed using a secure key-dependent hash. The hash is then XORed with a chosen binary logo and inserted into the LSBs of the same block. Verification proceeds in the reverse order first by calculating the key-dependent hash of the 7 MSBs in each block and XORing them with the LSBs. Comparison with the logo indicates tampered blocks. In the public key version, the 7 MSBs are hashed using a fixed hash, XORed with the logo and then encrypted using a public key encryption method. The encrypted bit-stream is again inserted in the LSBs of the same block. The verification algorithm proceeds by blocks and first calculates the hash of the 7 MSBs of all pixels in that block, XORs the hash with the decrypted LSBs (using the public key) and compares the result with the binary logo.

The logo can be either a binary picture with a graphical meaning or a randomly generated black and white pattern. If the logo has a visually recognizable structure, the tampered areas can be detected visually by comparison. Another advantage of using the logo is that cropping can be readily detected. The ability of this scheme to localize modifications is very satisfactory. The block size should be chosen so that the whole hash (128 bits) can be embedded. For example, block sizes of 8×16 or 12×12 pixels are possible.

Wolfgang and Delp scheme: This watermarking technique [6] consists in dividing the image into the blocks of about 64X64 pixels and inserting a robust mark into each block. To check the integrity of an image, the authenticator tests the presence or absence of the mark in all blocks. If the mark is present with a high probability in each block, then we can affirm that the image is authentic.

Like Van Schyndel et al. [1] , the authors recommend to use m-sequences [2] to generate the mark. The use of m-sequences is justified by the fact that they have excellent auto-correlation properties, as well as a very good robustness with noise addition. To generate the watermark, a binary sequence is mapped from {0, 1} to {-1.1}, arranged into a suitable block, and then added to the image pixel values.

Rey and Dugelay scheme: The basic idea of this method [7] consists in first extracting features from the original image and hiding them within a robust and invisible watermark. Then in order to check whether an image has been altered, we simply compare its features with those of the original image recovered from the watermark. If the features are identical, this will mean that the image is not tampered, otherwise the differences will indicate the altered areas.

The choice of image features used will directly affect the type of image alterations that we wish to detect. Additionally, those features will depend on the type of image under consideration. The features are typically selected so that invariant properties are maintained under weak image alterations and broken for malicious manipulations.

7. ANALYSIS

Our aim in this section is not to develop a list of all the possible malicious attacks that an image authentication system can be affected, but to show some of the frequent attacks which attacks the above four schemes. The analysis is shown in Table 1.

8. CONCLUSION

The increasing amount of digital exchangeable data generates new information security needs. Multimedia documents and specifically images are also affected. Users expect that robust solutions will ensure copyright protection and also guarantee the authenticity of multimedia documents.

In the current state of research, it is difficult to affirm which watermarking approach seems most suitable to ensure an integrity service adapted to images and more general way to multimedia documents. In this paper, we have only identified some attacks which are affecting some of the famous watermarking techniques. But how to protect valuable images and multimedia documents is also in the area of future research.

9. REFERENCES

[1] R. G. Van Schyndel, A. Z. Tirkel and C. F. Osborne, “A Digital Watermark”, in Proc IEEE International Conference on Image Processing, vol. 2, pp 86-90, Austin, Texas, USA, November 1994.

[2] J. G. Proakis, Digital communications, McGraw-Hill, New York, NY, USA, 3rd edition, 1995.

[3] S. P. Mohanty, “Digital watermarking: A tutorial review”, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.13.4913&rep=rep1&type=pdf (as on May 13, 2010).

[4] M. M. Yeung and F. Mintzer, “An invisible watermarking technique for image verification”, in Proc. IEEE International Conference on Image Processing, vol. 2, pp. 680-683, Santa Barbara, Calif, USA, October 1997.

[5] P. Wong, “A watermarking for image integrity and ownership verification”, in Proc. Final Program and Proceedings of the IS&-T PICS 99, pp. 374-379, Savana, Ga, USA, April 1999.

[6] R. B. Wolfgang and E. J. Delp, “A watermark for digital images”, in Proc. 1996 IEEE International Conference on Image Processing, vol. 3, pp. 219-222, Lausanne, Switzerland, September 1996.

[7] C. Rey and J.-L. Dugelay, “Blind detection of malicious alterations on still images using robust watermarks” in Secure Images and Image Authentication Colloquium, IEE Electronics & Communications, London, UK, 2000.

[8] 8. C. Rey and J.-L. Dugelay, “A Survey of Watermarking Algorithms for Image Authentication” in EURASHIP Journal on Applied Signal Processing 2002, pp. 613-621.