Int. J. Electron. Commun. (AEÜ) 70 (2016) 777–785

Contents lists available at ScienceDirect

International Journal of Electronics and Communications (AEÜ)

journal homepage: www.elsevier.com/locate/aeue

Cryptanalysis of an improved fragile watermarking scheme

Daniel Caragata[a,*], Juan Andres Mucarquer[a], Mirko Koscina[a], Safwan El Assad[b]

[a] Universidad Técnica Federico Santa María, Departamento de Electrónica, Avenida España 1680, Valparaíso, Chile.
[b] École Polytechnique de l'Université de Nantes, rue Christian Pauc, Nantes, France

ARTICLE INFO

Article history:
Received 30 November 2015
Accepted 2 March 2016

Keywords:
Fragile watermarking
Cryptanalysis
Markov chains
Multimedia security
Chaotic functions

ABSTRACT

This paper presents two attacks on Teng et al.'s fragile watermarking algorithm. Both attacks allow the attacker to apply valid watermarks on tampered images, therefore rendering the watermarking scheme useless. The first attack uses the watermarked version of two chosen images, and the second attack, a generalization of the first, uses a number of arbitrary watermarked images. The paper also models the cryptanalysis process for the second attack using Markov chains in order to demonstrate that the necessary number of images is relatively small for a high probability of successful attack. All the results that are presented in this paper have been confirmed by a practical implementation.

1. Introduction

In recent years we have witnessed the exponential growth of Internet and image sharing with different purposes, that vary from personal pictures to QR codes containing valuable pieces of information. For example, QR codes now hold information that can be used to enter a website, represent a verification checksum, digital signatures and public keys. Therefore, there is a need to search better ways to protect digital images against malicious manipulation or piracy. Digital watermarking is one of the techniques used for this end and consists in embedding information into digital data such that it is imperceptible to the human eye, while it can be easily and correctly detected by the watermarking algorithm [1].

Depending on the security purposes, watermarking algorithms can be classified as either robust, semi-fragile or fragile. Robust watermarking techniques are used for copyright protection and are resilient to a series of modifications made to the image, such as cropping, shrinking, rotating, etc. [2]. On the other hand, fragile watermarking is used to protect against malicious manipulation and is sensitive to any change made to the image. Semi-fragile schemes have the same purpose as fragile watermarking, but are tolerant to some image-processing operations made by a legitimate application [3] e.g. quantization noise from a lossy compression. Some examples of fragile and semi-fragile watermarking algorithms can be found in [4–7].

Teng et al. have cryptanalyzed a chaos-based fragile watermarking algorithm previously presented in [8], allowing an attacker to make undetected modifications on a watermarked image. Also, the authors proposed an improved version of the scheme that was supposed to be secure [9]. In this paper we present two attacks against Teng's et al. improved algorithm. These attacks do not recover the secret key. They recover secret information that allows the attacker to compute the same transformations or functions that are realized by the legitimate parties with the secret key. The first attack considers the scenario where an attacker has access to the watermarking device, and the second attack, a generalization of the first, only requires access to an arbitrary number of watermarked images. In both cases, an attacker would be able to apply valid watermarks on tampered images and, therefore would be able to circumvent the protection that the watermarking algorithm was supposed to offer. Moreover, we use Markov chains to demonstrate that the number of required images for the second attack is relatively small: less than 30 even for very large images.

The rest of the paper is organized as follows. Section 2 defines watermarking background, chaos-based watermarking schemes and a classification of attacks and vulnerabilities. Section 3 is a review of previous work on attacks to fragile watermarking schemes. Section 4 presents Teng et al.'s fragile watermarking algorithm, as well as the main design weaknesses that make the attacks possible. In Section 5 we present the two attacks, the results we have obtained implementing these attacks and we prove that the second attack does not require a large number of watermarked images to be successful. Section 6 concludes our work.

* Corresponding author. Tel.: +56 322654393.
E-mail addresses: daniel.caragata@usm.cl (D. Caragata), juan.mucarquer@usm.cl (J.A. Mucarquer), mirko.koscina@usm.cl (M. Koscina), safwan.el-assad@univ-nantes.fr (S. El Assad).

http://dx.doi.org/10.1016/j.aeue.2016.03.001
1434-8411/

2. Background of fragile watermarking schemes

A typical fragile watermarking scheme consists of two algorithms, watermark insertion and watermark verification. The insertion algorithm can be generalized by the following equation [10]:

$$I_w = E_k(I, W) \tag{1}$$

where I and I_w are the original and the watermarked image respectively, W represent the watermark information to be embedded, E_k is the insertion function under the user's secret key k.

The watermark verification process has two steps: watermark extraction and integrity check. Watermark extraction process can be generalized by the following equation [10]:

$$\widehat{W} = D_k(\widehat{I}) \tag{2}$$

where $\widehat{W}$ represents the extracted watermark information of the possibly corrupted image $\widehat{I}$. D_k is the watermark extraction function.

Integrity check process generally verifies if the extracted watermark fulfills a certain condition, e.g. being identical to an expected value. Furthermore, the verification process can give additional information about the type of distortion the image $\widehat{I}$ had suffered, such as the region that has been corrupted.

2.1. A classification of vulnerabilities and attacks to watermarking schemes

In general, attacks on fragile watermarking schemes aim to break the integrity verification function, i.e. they allow the attacker to make undetected modifications to watermarked images.

Common security problems of watermarking schemes, classified in [11]:

1. The ability of an attacker to make *undetected modifications* to the watermarked image: a successful attack should modify the image in such a way that there is a reasonable probability it goes undetected by the watermarking scheme. For an attack to fall under this category, the attacker does not need to recover any secret information.
2. The discovery of *information leaks* from watermarked images: the attacker is able to recover information about the key or equivalent secret information, which he later uses to make undetected modifications.
3. *Protocol weaknesses*: vulnerabilities that are not part of the watermarking scheme itself, but exploit the practical implementation of the protocol. For example, the protocol may allow an attacker unlimited access to the verification device (oracle attack).

According to [11] and based on the information and capabilities available to the attacker, the may attacks fall into one of the following five categories:

1. *Stego-image attack*: The attacker only possesses one authenticated image; it is similar to cryptographic ciphertext only attack.
2. *Multiple stego-image attack*: The attacker has multiple watermarked images; also similar to ciphertext only attack, but with more ciphertext.
3. *Verification device attack*: The attacker is capable of verifying the integrity of any image: it is similar to chosen ciphertext attack.
4. *Cover image attack*: The attacker has a number of pairs of original-watermarked images; it is similar to known plaintext attack.
5. *Chosen cover-image attack*: The attacker has access to the watermarking device and can apply the watermark to images he chooses; it is similar to cryptographic chosen plaintext attack.

2.2. Chaos theory and watermarking

The main characteristic of chaotic systems is their very high sensitivity to initial conditions. It has been claimed that the idea of using this property in cryptography dates back to Shannon's 1952 seminal paper [12].

Even simple mathematical functions show chaotic behavior [13] and can be used as cryptographic primitives [12] to construct versatile encryption schemes. One classical example of a simple chaotic map used in cryptography is a two-degree polynomial difference equation, called the Logistic map:

$$x_{k+1} = \mu x_k(1 - x_k) \tag{3}$$

This non-linear recurrence equation is in chaotic state when $3.57 < \mu \leq 4$.

Arnold's Cat Map [14] is another chaotic map that has been widely adopted in image cryptography as a permutation component [8,15–17]. It is a two-dimensional invertible map described by the following linear transformation:

$$\begin{bmatrix} x_{n+1} \\ y_{n+1} \end{bmatrix} = \begin{bmatrix} 1 & a \\ b & ab+1 \end{bmatrix} \begin{bmatrix} x_n \\ y_n \end{bmatrix} \mod N = A \begin{bmatrix} X_n \\ Y_n \end{bmatrix} \mod N \tag{4}$$

where a, b are positive integers and $(x_n, y_n) \in \{0, 1, \ldots, N-1\} \times \{0, 1, \ldots, N-1\}$. This is a periodic transformation, i.e. the pixel at the position (x, y) returns to the original position after T iterations, where T is a function of a, b and N. The integers a, b and the number of times the map is applied are typically secret parameters.

3. Related work

This section presents relevant previous works of cryptanalysis and improvement on fragile watermarking algorithms. We will briefly describe the schemes, classify the security flaws that made the attacks possible and describe the attacks. We refer the reader to the original works for more detailed information regarding the implementation and performance of the attacks.

Rawat et al. [8] proposed an image authentication scheme that was cryptanalyzed and improved, firstly by Teng et al. [9] (whose security is analyzed and broken later in this paper) and secondly by Botta et al. [18], that has also been recently attacked in [19].

In order to apply a watermark W to a $m \times n$ image, Rawat et al.'s scheme first applies Arnold's Cat Map k times on the image and then generates a chaotic sequence of length $m \times n$ using the initial condition as secret key. Each element of the chaotic sequence is rounded off to a binary value, which is then XORed with the watermark image. The result of this operation is the secret watermark, W_p, that will be inserted in the LSB plane of the image. Finally, Arnold's cat map is applied $T - k$ times in order to obtain the original image, where T is the Arnold's Cat Map period.

The main flaw of Rawat et al.'s algorithm is that the secret watermark W_p depends only on the secret key and on the watermark image, W. Furthermore, the attacker knows where the watermark has been inserted because the scheme uses the LSB of every pixel.

The scheme is vulnerable to *undetected modifications*: the attacker saves the LSB plane of the watermarked image and modifies the other 7 bit planes of the image. A countermeasure proposed by Botta was to craft a more complex function for calculating the watermark bit, that includes the MSBs and the pixel position (x, y), so that modifying the MSBs will affect the watermark bit. Therefore, this attack that simply keeps the LSB plane unchanged and modifies the 7 MSB planes should be detected in the extraction process.