# Invertible chaotic fragile watermarking for robust image authentication

Panagiotis Sidiropoulos, Nikos Nikolaidis and Ioannis Pitas*

Department of Informatics, Aristotle University of Thessaloniki

Box 451, Thessaloniki 54124, GREECE

psid@iti.gr, {nikolaid, pitas}@aiia.csd.auth.gr

December 29, 2008

## Abstract

Fragile watermarking is a popular method for image authentication. In such schemes, a fragile signal that is sensitive to manipulations is embedded in the image, so that it becomes undetectable after any modification of the original work. Most algorithms focus either on the ability to retrieve the original work after watermark detection (invertibility) or on detecting which image parts have

*This work has been partially supported by the European Commission through the IST Programme under Contract IST-2002-507932 ECRYPT.

been altered (localization). Furthermore, the majority of fragile watermarking schemes suffer from robustness flaws. We propose a new technique that combines localization and invertibility. Moreover, watermark dependency on the original image and the nonlinear watermark embedding procedure guarantees that no malicious attacks will manage to create information leaks.

# 1 Introduction

In the last decades substantial advantages in multimedia and web technology have facilitated the production and distribution of image content. However, the ease of media copying and editing also enables unauthorized use and tampering of the media content. One of the applications of digital watermarking is multimedia authentication and content integrity verification. Watermarking schemes for multimedia authentication try to detect forgeries, i.e. images that have been modified (intentionally or not).

Fragile [1]–[10] and semi-fragile [18]–[21] watermarks are usually employed for image authentication. Semi-fragile watermarks manage to identify any malicious image tampering while being tolerable to jpeg compression or slight image content alterations. However, there are applications where even the slightest modification indicates a content degradation and needs to be identified. In these cases the implementation of a total fragile watermarking technique is required. Such watermarks are expected to be sensitive both to malicious attacks and to incidental content manipulations [1].

There are several prerequisites for an efficient fragile watermarking technique: a) the fragile watermark must be perceptually transparent, b) tampering should be detected without using the original image, c) despite being fragile, the technique must be robust to malicious attacks that try to sabotage the watermark functionality, d) the technique must be able to locate the tampered regions within an image (localization).

The majority of existing fragile watermarking algorithms suffer from robustness flaws. For example, many pixel-wise techniques are vulnerable to oracle attacks, i.e., attacks in which the forger has unlimited access to the detector and tries with contiguous tests to create a tampered image that is still detected as authentic [4].

Furthermore, most fragile watermarking techniques change permanently the image, i.e. they are not invertible. In this case, the watermark that is embedded in an image cannot be removed. Thus, the watermarking procedure corrupts the image content to a degree that depends on the watermark signal power. In some applications, such as medical imaging, even a minimal image distortion might be unacceptable. This led to the development of invertible (or erasable) watermarking techniques for authentication purposes [2]. Unfortunately the robustness of such techniques to attacks is rather mediocre.

A novel invertible fragile watermarking technique for robust image authentication is proposed in this paper. The watermark is generated by a pseudo-random chaotic process that involves the values of the original image pixels. Thus, it is image content dependent. In order to extract an existing watermark in a water-

marked image, the exact knowledge of system parameters is required. Furthermore, if the image is not modified, the watermark can be completely removed from the original image. An attacked (edited) version of the image can be detected with almost 100% probability. Furthermore, the tampered image region can be well localized in the majority of the tampering cases. Additionally, extreme sensitivity of the chaotic function with respect to the initial conditions ensures that an approximate knowledge of the system parameters will not reveal the watermark. The watermark dependency on the image content strengthens the system robustness to attacks that seek to expose the watermark signal form and, finally, harm the functionality of the watermarking system.

The structure of this paper is organized as follows. Section 2 describes the proposed scheme. Important aspects such as chaotic synchronization, localization and security are further analyzed in Section 3. Experimental results that demonstrate the fragile watermarking scheme performance are presented in Section 4. Conclusions follow in Section 5.

# 2 Method Description

## 2.1 Watermark Embedding

The watermark generation process is applied on the spatial image domain, while scanning an image in a row-wise manner. In the simplest version, starting from the top left image corner, we evaluate iteratively the variable $X$ as follows:

$$X_i = f(X_{i-1}) + m \cdot Y_i \tag{1}$$

where $f$ is a non-linear function, $Y_i$ is the luminance of i-th image pixel and $m$ is a scaling factor. From the above equation and, since the previous value $X_{i-1}$ is used to evaluate $f$ at pixel position i, it is obvious that, if the values of the non-linear function $f$ range from $f_{min}$ to $f_{max}$ , then the range of $X_i$ should be the interval $[f_{min}+m\cdot Y_{min}, f_{max}+m\cdot Y_{max}]$ . For example, if the image intensity $Y$ is represented by 8-bit values, then this interval becomes $[f_{min}, f_{max}+255\cdot m]$ . Thus, we can conclude that certain non-linear functions that have been used in the past for watermark generation, such as skew tent map, Bernoulli shift, logistic map etc [11]–[13], [17] cannot be used in the watermark generation procedure, since they have identical domain and range of values. Instead, in this work, a non-linear function with unbounded domain is used, namely the chaotic Chebyshev function [14] that is iteratively computed by the following formula:

$$f_{CHEB}(X_{i-1}) = \tanh(C_1 \cdot X_{i-1}) - b \cdot \tanh(C_2 \cdot X_{i-1}) \tag{2}$$

Hyperbolic tangent is an increasing function that takes values in the range $[-1, 1]$ . Consequently, if we assume, without loss of generality, that the constants $b$, $C_1$ and $C_2$ are positive and $C_1 > C_2$ then:

$$If \;\; X > 0, \;\; 1 > \tanh(C_1 X) > \tanh(C_2 X) > 0 \Longrightarrow f_{CHEB} < \tanh(C_1 X) < 1$$

$$If \;\; X < 0, \;\; 0 > \tanh(C_2 X) > \tanh(C_1 X) > -1 \Longrightarrow f_{CHEB} > \tanh(C_1 X) > -1$$

The parameter $b$ is arbitrarily chosen to be equal to 1.6 [14]. For this value, $f$ generates values in the interval $[-1, 1]$ . In order to have a non-linear function $f$ with a pseudorandom behavior, it is essential that the $f$ values are uniformly distributed in the region $[-1, 1]$. The selection of the other system parameters should be made so as to fulfill this condition. We can easily deduce, that if $C_1 > C_2 \gg 1$ then for most $X$, $\tanh(C_1 \cdot X) \approx \tanh(C_2 \cdot X) \approx \pm 1$ and $f_{CHEB}(X) \approx \pm(b-1)$ , where the sign of $f$ depends on the sign of the previous value of $X$. On the other hand, if $C_1 \approx C_2$ then $f_{CHEB} \approx \tanh(C_1 \cdot X) \cdot (1-b)$ and consequently $f$ maps its domain to the smaller interval $[1-b, b-1]$. Thus, in order to guarantee that the Chebyshev function will map its domain strictly and as uniformly as possible in the range $[-1, 1]$ , we have selected $C_1 \gg C_2 > 0$. In our implementation, we chose $C_1 = 200$ and $C_2 = 2$ respectively. Experiments on the pdf of $f_{CHEB}$ output are described in section 3.2.

After the calculation of the pseudo-random signal $X$ using (1), the binary watermark $W$ is generated by quantizing $X$ with the sign function. $W$ is then additively embedded on the original image, thus leading to the watermarked image $Y_W$ :

$$W_i = \text{sgn}(X_i) \tag{3}$$

$$Y_{W_i} = Y_i + W_i \tag{4}$$

## 2.2 Watermark Detection

The initial value $X_1$ and/or the exact system parameter values can be used as the watermark key. Watermark detection is performed in a blind fashion, i.e., without the need to resort to the original image. Starting from $X_1$, we reverse the procedure based on the fact that since equations (1), (3) and (4) hold, $X$ must satisfy one of two specific inequalities. More specifically, by using (4), (1) can be rewritten as:

$$X_i = f(X_{i-1}) + m \cdot (Y_{W_i} - W_i) \tag{5}$$

If we denote with $A_i$ the quantity $f(X_{i-1}) + m \cdot Y_{W_i}$, the above equation takes the following form:

$$X_i = A_i - m \cdot \text{sgn}(X_i) \tag{6}$$

For every pixel, the value and sign of $X$ must not contradict each other, i.e. if $X_i > 0$ then $\text{sgn}(X_i) = 1$ and if $X_i < 0, \text{sgn}(X_i) = -1$ . This property leads to a system of inequalities. More analytically, if:

$$X_i > 0, sgn(X_i) = 1 \Rightarrow X_i = A_i - m \cdot sgn(X_i) = A_i - m > 0 \Rightarrow A_i > m$$

$$X_i < 0, sgn(X_i) = -1 \Rightarrow X_i = A_i - m \cdot sgn(X_i) = A_i + m < 0 \Rightarrow A_i < -m$$

Consequently, if the detection parameters are the same as the ones used for embedding and the image is not tampered, the value of $A_i$ will never reside in the interval $[-m, m]$ . This property is used for the detection of image tampering. On the other hand, if the image is authentic the property can be used for watermark removal and recovery of the original image. The following algorithm can be used for watermark detection.

Starting from the second pixel and scanning the image in a row-wise manner:

1. $A_i = f(X_{i-1}) + m \cdot Y_{w_i}$
2. If $A_i > m$ then $W_i = 1$ and $X_i = A_i - m$
3. If $A_i < -m$ then $W_i = -1$ and $X_i = A_i + m$

If $-m < A_i < m$ holds even for one image pixel, then either an invalid detection key has been used or an image tampering has been identified. The process is immediately terminated and the image is classified as non-authentic. If the detection process finishes and no values of $A_i$ were found within the $[-m, m]$ range, then the image is classified as authentic and the original image can be recovered by subtracting the evaluated watermark $W$ from the authenticated image.

# 3 Algorithm Analysis

## 3.1 Chaotic synchronization

Non-linearity in watermark embedding allows us to exploit the properties of chaos, like its extreme sensitivity to initial conditions, which ensures that a watermark generated even by a key in the vicinity of the correct one will not be positively (and erroneously) detected in a watermarked image, since the two chaotic trajectories will rapidly diverge. Unfortunately, chaos also causes the side-effect of synchronization. In 1990, Pecora and Caroll found that, under certain circumstances, two chaotic systems that are linked with a common signal or signals synchronize, i.e. the trajectory of one system will converge to that of the other system and, from that point onwards, they will remain synchronized [15]. Later, Maritan and Banavar expanded the previous result in the case where chaotic systems are linked by the same type of white or pink $(1/f)$ noise [16]. They also proved that, if the commonly present noise, is stronger than a certain level (approximately 0 dB), the trajectories for various initial conditions become point by point identical after a certain time that depends on the noise power.
In our case, any image can be fairly well modelled as a combination of white and pink noise. From this point of view, equation (1) represents a non-linear iterative function $f_{CHEB}$ that is exposed to such a white and pink noise, namely the image intensity signal $Y_i$ . If the image power $Y_i$ in (1) is higher than a certain level, the correct watermark and a watermark that is generated by a false key synchronize, since they are linked by the same noise (the image),

and converge rapidly to the same signal. In this situation, the space spanned by the watermarks degenerates to only a few signals per image, and the probability of false acceptance becomes prohibitively high, since two watermarks generated by different keys and parameters and applied on the same image will converge to the same values and, thus, their separation during detection will be usually impossible. From the above, it is easily deduced that, in order to circumvent chaotic synchronization, the image intensity must be scaled by a factor $m$. Figure 1 depicts the mean dissimilarity of pairs of different watermarks $D = \frac{1}{N}\sum_{i=1}^{N}|W_1(i) - W_2(i)|$ that are generated for a test image by randomly chosen keys versus the value of the factor $1/m$. The Figure shows that the image intensity should be scaled down by at most a factor $m = 1/300$ (for 8 bit image pixels) in order to produce a system that is not vulnerable to chaotic synchronization. Indeed, it can be seen from Figure 1 that for $m$ bigger than $\sim 1/300$ the distinct watermarks are identical in more than 95% of the image pixels and for m bigger than $\sim 1/250$ any watermark key will produce almost the same signal.

### 3.2 Tampered Region Localization

It is very desirable for a fragile watermarking system to be able to localize the tampered image regions. Furthermore, in many occasions, a conclusion regarding the tampering goals and/or type could be derived from the semantic content of the original areas and the localization of the tampered regions [2],[3],[9]. The extreme sensitivity of the proposed scheme to image modifications can be used

to achieve an adequate tamper localization.

As already mentioned, a tampered image (or use of the wrong watermark key) will be declared whenever the quantity $A_i$ falls in the range $[-m, m]$. $A_i$ consist of two parts that can be assumed independent, the non-linear function value $f_{CHEB}(\cdot)$ and the normalized image intensity $m \cdot Y_{W_i}$ . This assumption is only partially correct, because the two parts are actually correlated since they come from two adjacent image pixels. However it was deemed necessary in order to make the analysis tractable. Moreover, its adoption resulted in theoretical results that are sufficiently close to the experimental ones, as will be shown at the end of this section.

If the embedding parameters are selected according to the previous analysis, it was proven experimentally that approximately one third of $f_{CHEB}(\cdot)$ values spreaded uniformly in the interval $[-1, 1]$ and the rest two thirds were evenly distributed in the vicinities of $(1 - b)$ and $(b - 1)$. Thus, we can model the probability density function of $f_{CHEB}(\cdot)$ output as a combination of a uniformly distributed function plus two pulses of equal height at positions $(1 - b) = -0.6$ and $(b-1) = 0.6$. Furthermore the image histogram is assumed (for making the analysis tractable) to be continuous and uniform. Thus, the probability that an image tampering is detected in a specific pixel is equal to the probability that $A_i$, i.e., the sum of the two random variables $f_{CHEB}(\cdot) + m \cdot Y_{W_i}$ will fall in the interval $[-m, m]$. Since the two parts of the sum are considered to be independent, the density function of quantity $A_i$ equals to the convolution of the densities of these two parts. Due to chaotic synchronization, $m$ is restricted

to be less than 1/256 for 8 bit images, as described in the previous section. If $Y_i' = m \cdot Y_{W_i}$ is the scaled luminance of the i-th pixel and $f_Y$ the probability density of the luminance, then:

$$f_Y(y) = \frac{1}{255 \cdot m}[u(y) - u(y - 255 \cdot m)]$$

where $u()$ denotes the unit step function. Furthermore, if we denote with $f_X$ the probability density of the Chebyshev function then:

$$f_X(x) = \frac{1}{6}[u(x+1)-u(x-1)]+\frac{1}{3}\delta(x-1+b)+\frac{1}{3}\delta(x+1-b) = f_{X1}(x)+f_{X2}(x)+f_{X3}(x)$$

The density function of $A_i$, $f_A$ is evaluated by the following formula:

$$f_A(\alpha) = f_X \star f_Y = \int_{-\infty}^{\infty} f_{X1}(\alpha - y)f_Y(y)dy + \int_{-\infty}^{\infty} f_{X2}(\alpha - y)f_Y(y)dy + \int_{-\infty}^{\infty} f_{X3}(\alpha - y)f_Y(y)dy = f_{A1} + f_{A2} + f_{A3}$$

It can be easily proven that $f_{A1}$ is given by the following formula:

$$f_{A1}(\alpha) = \begin{cases} 0 & \text{if } \alpha < -1 \\ \frac{\alpha+1}{6 \cdot 255 \cdot m} & \text{if } -1 < \alpha < -1 + 255 \cdot m \\ \frac{1}{6} & \text{if } -1 + 255 \cdot m < \alpha < 1 \\ \frac{255 \cdot m - \alpha + 1}{6 \cdot 255 \cdot m} & \text{if } 1 < \alpha < 1 + 255 \cdot m \\ 0 & \text{if } \alpha > 1 + 255 \cdot m \end{cases}$$

In a similar way, the two other parts of this density function $f_{A2}$ and $f_{A3}$ are given by the following formulas:

$$f_{A2}(\alpha) = \begin{cases} 0 & \text{if } \alpha < 1-b \\ \frac{1}{3 \cdot 255 \cdot m} & \text{if } (1-b) < \alpha < (1-b+255 \cdot m) \\ 0 & \text{if } \alpha > (1-b+255 \cdot m) \end{cases}$$

$$f_{A3}(\alpha) = \begin{cases} 0 & \text{if } \alpha < b-1 \\ \frac{1}{3 \cdot 255 \cdot m} & \text{if } (b-1) < \alpha < (b-1+255 \cdot m) \\ 0 & \text{if } \alpha > (b-1+255 \cdot m) \end{cases}$$

The probability $P$ of tampering detection at a certain pixel location, namely the probability that the value of $A_i$ falls inside the range $[-m, m]$ for a certain pixel of a manipulated image is:

$$P = \int_{-m}^{m} f_A(\alpha)d\alpha = \int_{-m}^{m} (f_{A1}(\alpha) + f_{A2}(\alpha) + f_{A3}(\alpha))d\alpha = \frac{m}{3} + \int_{-m}^{m} f_{A2}(\alpha)d\alpha + \int_{-m}^{m} f_{A3}(\alpha)d\alpha$$

In our case $m < b-1$ , since $m < \frac{1}{256}$ and in our implementation $b = 1.6$. For such values the previous integrals can be proven to be:

$$\int_{-m}^{m} f_{A2}(\alpha)d\alpha = \begin{cases} \frac{2}{765} & \text{if } m > \frac{b-1}{254} \\ \frac{256 \cdot m+1-b}{255 \cdot m} & \text{if } \frac{b-1}{256} < m < \frac{b-1}{254} \\ 0 & \text{if } m < \frac{b-1}{256} \end{cases}$$

and

$$\int_{-m}^{m} f_{A3}(\alpha)d\alpha = 0$$

Thus, the probability $P$ that image tampering is detected at a certain pixel location is:

$$P = \begin{cases} \frac{m}{3} & \text{if } m < \frac{1}{426.6} \\ \frac{m}{3} + \frac{256 \cdot m - 0.6}{255 \cdot m} & \text{if } \frac{1}{426.6} < m < \frac{1}{423.3} \\ \frac{m}{3} + \frac{2}{765} & \text{if } m \geq \frac{1}{423.3} \end{cases}$$

It is known that an event whose probability of occurrence is $\hat{P}$ will be expected to happen for the first time after $\frac{1}{\hat{P}}$ trials. Consequently, it is deduced that any manipulation of a watermarked image is expected to be detected after scanning $(\frac{1}{P})$ pixels after the first tampered pixel encounter (in a row-wise manner). In other words, if a certain image pixel is tampered the detection will signal a tampering $(\frac{1}{P})$ pixels after this pixel. The experimentally selected scaling factor $m$ varies between 1/300 and 1/400. For these values, tamper localization is expected to happen after scanning 268 to 290 pixels after the first encountered tampered pixel. Experimental tamper position localization is almost 25% better than the one provided by the theoretical analysis.

Tampering localization accuracy can be dramatically enhanced by embedding more than one watermarks in a pyramidal way. Firstly, the whole image is watermarked using a single watermark. Then, the image is divided into non-overlapping blocks of $W \times H$ pixels and a different watermark is embedded in each block. The watermarked image is subsequently divided again into disjoint

blocks of $\frac{W}{2} \times \frac{H}{2}$ pixels and the new blocks are independently watermarked again. In every step of the procedure a new layer of watermarks is superimposed, leading to a major improvement of systems localization at the expense of perceptual quality.

Since every watermark is "unique", due to the fact that its values depend on the underlying content, watermarks in every block or layer can be produced using the same key. Alternatively, keys for the various watermarks can be produced by values generated by a non-linear function (for example a skew-tent map that is initialized with one key) Thus, only one initial value-key is required to generate all layers of watermarks.

During detection, the watermarks are detected in each layer starting from the upper layer and subsequently erased from the image before proceeding to the detection on the next (lower) layer. An image is characterized as authentic, if all watermarks are detected. The pyramidal scheme facilitates localization of an image modification, as any tampering detection in a certain block, will be triggered from a tampering of the specific block content. The previously estimated values of localization (i.e. the fact that tamper localization is expected to happen scanning 268 to 290 pixels after the first encountered tampered pixel) allow us to confine modifications in blocks of $32 \times 32$ pixels and, in many cases, even in blocks of $16 \times 16$ or $8 \times 8$ pixels. The watermark that is embedded in the entire image ensures that an alteration that will fail to be localized will be finally detected. Additionally, the watermark that is embedded in the entire image can be used to distinguish a patchwork of different or not authenticated

images from an actually authentic image.

In addition, the probabilities of false positive and false negative detection, which characterize the performance of a watermarking scheme, can be evaluated for the proposed method. In a fragile watermarking scheme, false positive detection probability is defined as the probability that a tampered image is identified as authentic. From the previous analysis, the theoretical value of the false positive detection probability $P_f$ for an image of dimensions $M \times N$ can be easily shown to be:

$$P_f = (1 - P)^{M \cdot N} \tag{7}$$

If $\frac{1}{300} < m < \frac{1}{400}$ , the false positive probability detection $P_f$ is between $10^{-25}$ and $10^{-27}$, even for a small $128 \times 128$ image. Thus, it is easily concluded that the probability of false positive detection is practically zero. Moreover, the watermark embedding procedure is chaotic and, thus, deterministic. Consequently, the probability of an authentic image to be identified as tampered, i.e. the false negative detection probability, is by definition equal to zero.

## 3.3 Robustness Considerations

From the previous discussion, it can be concluded that the proposed technique will, in general, detect any image tampering. In this section, we will further analyze and discuss the robustness and other properties of the proposed fragile watermarking technique. At first, it must be noted that attacks on fragile watermarking schemes differ significantly from attacks on robust watermarking

techniques. In fragile watermarking, the attacker is not interested in making the watermark undetectable. Actually, destroying the watermark is quite easy, because of its inherent fragility. Attacks to fragile watermarking schemes aim to extract the watermark pattern in order to use it to illegally authenticate images, replace pre-existing watermarks with fake ones and tamper an authentic image in a way that the modification will not be detected. There are several levels of attacks, depending on the information and devices available to the attacker. As will be explained below, the proposed watermark technique is expected to be robust to many different kinds of known attacks, such as those described in [3]–[10].

The robustness of our scheme stems from the fact that the values of the watermark are determined by at least three factors: the image, the parameters of the chaotic function and the initial value of the chaotic watermark generation function. Changes to any of these factors will lead to a substantially different watermark signal. Content dependency in particular, provides robustness against all attacks that are based on the possession of multiple pairs of original-watermarked images [3],[4],[8], since the watermark that is superimposed to an image is unique and defined by its content. Furthermore, the non-linearity of the watermark generation procedure deters any information leakage from knowledge of a specific watermark.

Actually, the only way to attack the proposed method using approaches that are based to an unlimited access to the watermark detection algorithm is through a brute-force attack. For example, watermark detection attacks that use the min-

imum undetectable watermark modifications in order to derive a manipulated image that is classified as authentic [3],[4],[8],[10] pose no threat to the proposed scheme, since the slightest change of the image content will be detected. Even if an attacker has unlimited access to the watermark embedding algorithm (chosen cover image attack) [3], he can not extract the watermark pattern, if all the chaotic function parameters are not accurately known. One might think that an easy way to retrieve these parameters is by selecting arbitrary parameter values and comparing the generated watermark with the original one. If the watermarks are similar enough, this would mean that the parameter values used for the generation of the two watermarks are very close. For such an attack to be successful, the watermark dissimilarity must be a monotonically increasing function of the parameter distance, which is not the case in chaotic systems. This fact has been also experimentally verified for the proposed system, by evaluating the similarity of watermarks generated by parameters whose values are converging.

# 4 Experimental Results

As mentioned previously, the watermarks can be reconstructed and removed during watermark detection. As a consequence, the proposed method is fully invertible and, thus, invertibility need not be verified by simulation. However, a series of other experiments were conducted. The first set of these experiments dealt with the evaluation of the false positive rate of the proposed method. For

this purpose, we watermarked various images with 100,000 randomly generated watermarks and for each of them we performed detection with a watermark that was generated by a key that differed less than $10^{-6}$ from the correct one, which is equivalent to trying to detect a watermark on a tampered image. The simulation demonstrated that no false positive detection was encountered in any of the performed tests.

Next, we tested the pyramidal scheme by superimposing 4 layers of watermarks on various images of size $256 \times 256$ , one in the whole image, and 3 by splitting the image in blocks of $32 \times 32$, $16 \times 16$ and $8 \times 8$ pixels respectively, in a pyramidal way. To evaluate the watermark's localization potential, we randomly changed the intensity value in just one randomly chosen pixel per block. The experiments involved 100,000 watermarks and correct tamper detection (i.e. characterization of the image as being tampered) was achieved from the proposed system (4 layers) in all cases. Furthermore, the 3 block-based, tamper localization layers achieved correct tamper detection (and thus good localization) in 99.95% of the cases. Additionally, we interchanged two $32 \times 32$ blocks of the watermarked image and observed that the corresponding image was identified as a collage of authentic blocks.

Finally we produced semantically altered versions of watermarked images and tested the system's ability to identify and localize malicious content modifications. Figures 2a, 3a and 2b, 3b show authentic images and their watermarked versions. Figures 2c, 3c, depict the modified images. The image in Figure 2 was exposed to additive noise whereas that in Figure 3 has been altered by re-

moving the tree branches in the upper right corner. Our pyramidal watermark detector recognized the images in Figures 2c, 3c as non-authentic and localized the tampering in the area where it actually happened. The same results were also obtained in other images with similar modifications.

# 5 Conclusions

In this work, we proposed a novel invertible fragile watermarking technique that can detect and localize manipulations of an image. We have proven that the false positive and false negative detection probabilities of the proposed algorithm are practically zero. Watermarks generated by the algorithm can be easily and fully removed from the authenticated image. A pyramidal version of the technique can exploit the extreme sensitivity of the system in order to achieve very good localization. Also, the scheme exploits chaos properties and content dependency to achieve robustness to known fragile watermarking attacks that try to sabotage its functionality. A theoretical analysis of the proposed method has been performed whereas experimental results were very satisfactory.

# References

[1] Katzenbeisser S. and Petitcolas F. Information Hiding Techniques for Steganography and Digital Watermaking, Artech House Inc., 2000

[2] Cox I. J., Miller M.L. and Bloom J. A. Digital watermarking, Morgan Kaufman Publishers, 2002

[3] Lin E. and Delp E. A review of fragile image watermarks, Proc. of the Multimedia and Security Workshop (ACM Multimedia'99), Orlando, 1999, pp. 25-29

[4] Fridrich J. Security of fragile authentication watermarks with localization, SPIE Photonic West, San Jose, 2002, pp. 691-700

[5] Yeung M. M. and Mintzer F. C. An invisible watermarking technique for image verification, Proc. IEEE International Conference on Image Processing (ICIP) 1997, vol. 2, pp. 680-683

[6] Wong P. A watermark for image integrity and ownership verification, Proc. IS & T PIC Conference, Portland, Oregon, 1998

[7] Si H. and Li C. T. Fragile watermarking scheme based on the block-wise dependence in the wavelet domain, Proc. of the multimedia and security workshop (ACM Multimedia 04), Magdenburg, 2004, pp. 214-219

[8] Holliman M. and Memon N. Counterfeiting attacks on oblivious block-wise independent invisible watermarking schemes, Proc. IEEE Trans. on Image Processing, 2000, vol. 9, no. 3, pp. 432-441

[9] Fridrich J., Goljan M. and Du R. Invertible authentication, Proc. SPIE, Security and watermarking of multimedia contents, San Jose, 2001

[10] Fridrich J. et al. Further Attacks on Yeung-Mintzer Fragile Watermarking Scheme, Proceedings of IS&T/SPIE's 12th Annual Symposium, Electronic

Imaging 2000: Security and Watermarking of Multimedia Content II, volume 3971, San Jose, 2000

[11] Tefas A., Nikolaidis N. and Pitas I. Chaotic watermark sequences for correlation-based schemes, Proc. of 12th European Signal Processing Conf. (EUSIPCO 2004), Vienna, Austria, 2004, pp. 1891-1894

[12] Mooney A., Keating J. G. and Pitas I. A Comparative Study of Chaotic and White Noise Signals in Digital Watermarking, Chaos, Solitons and Fractals, 2008, vol. 35, no. 5, pp 913-921

[13] Tefas A., Nikolaidis A., Nikolaidis N., Solachidis V., Tsekeridou S. and Pitas I. Markov chaotic sequences for correlation based watermarking schemes, Chaos, Solitons and Fractals, 2003, vol. 17, no. 2, pp. 567-573

[14] Lai C.-H. and Zhou C. Synchronization of chaotic maps by symmetric common noise, Europhys. Letters, 1998, vol. 43, pp. 376-379

[15] Pecora L. M. and Carroll T. L. Synchronization in chaotic systems, Phys. Review Letters, 1990, vol. 64, no. 8, pp. 821-824

[16] Maritan A. and Banavar J. R. Chaos, Noise and Synchronization, Phys. Review Letters, 1994, vol. 72, no. 10, pp. 1451-1454

[17] Mooney A., Keating J. G. and Heffernan D. M. A Detailed Study of the Generation of Optically Detectable Watermarks using the Logistic Map, Chaos, Solitons and Fractals, 2006, vol. 30, no. 5, pp. 1008-1097

[18] Peng Z, Liu W. Color image authentication based on spatiotemporal chaos and SVD, Chaos, Solitons and Fractals, 2008, vol. 36, no. 4, pp 946-952

[19] Zou D., Shi Y.Q., Ni Z., Su W. A Semi-Fragile Lossless Digital Watermarking Scheme Based on Integer Wavelet Transform, IEEE Transactions on Cirucits and Systems for Video Technology, 2006, vol. 16, no. 10, pp. 1294-1300

[20] Lin E. T., Podilchuk C. I., and Delp E. J. Detection of image alterations using semi-fragile watermarks, Proc. of IS&T/SPIE's 12th Annual Symposium, Electronic Imaging 2000: Security and Watermarking of Multimedia Content II, volume 3971, San Jose, 2000

[21] Boyer J.P., Duhamel P., Blanc-Talon J. Game-Theoretic Analysis of a Semi-Fragile Watermarking Scheme Based on SCS, Proc. IEEE International Conference on Image Processing (ICIP) 2005, vol. 2, pp. 1122-1125

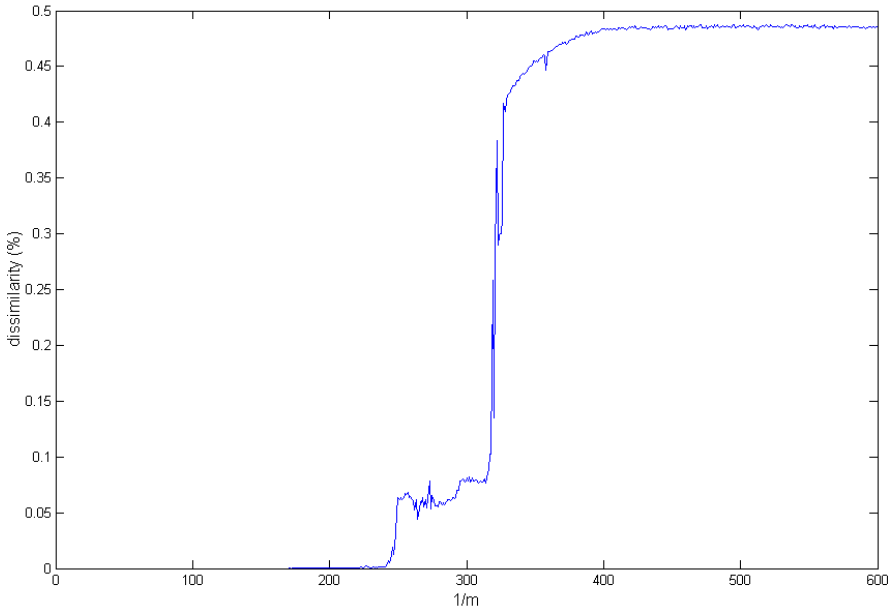


Figure 1: Watermark distance D for watermarks generated by different keys for the same image. The image intensity is scaled by a factor of $m$ before being added to $f_{CHEB}$.

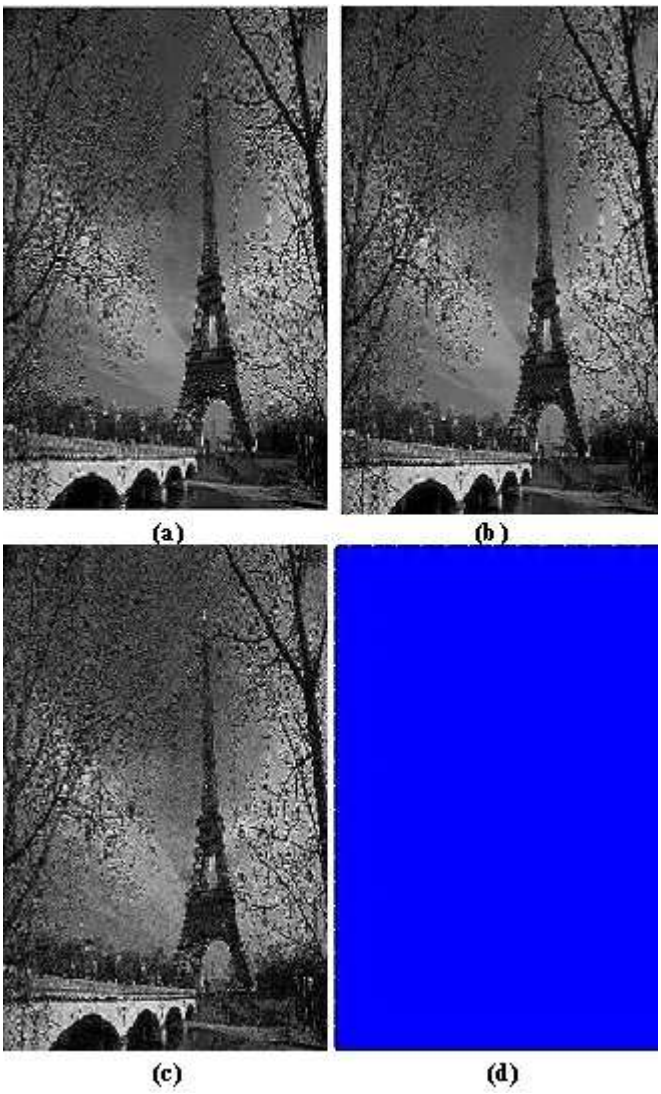


Figure 2: (a) Original Image, (b) Watermarked version, (c) Tampered version, exposed to additive noise, (d) Detection results: the grey area is considered authentic while blue areas are considered tampered.

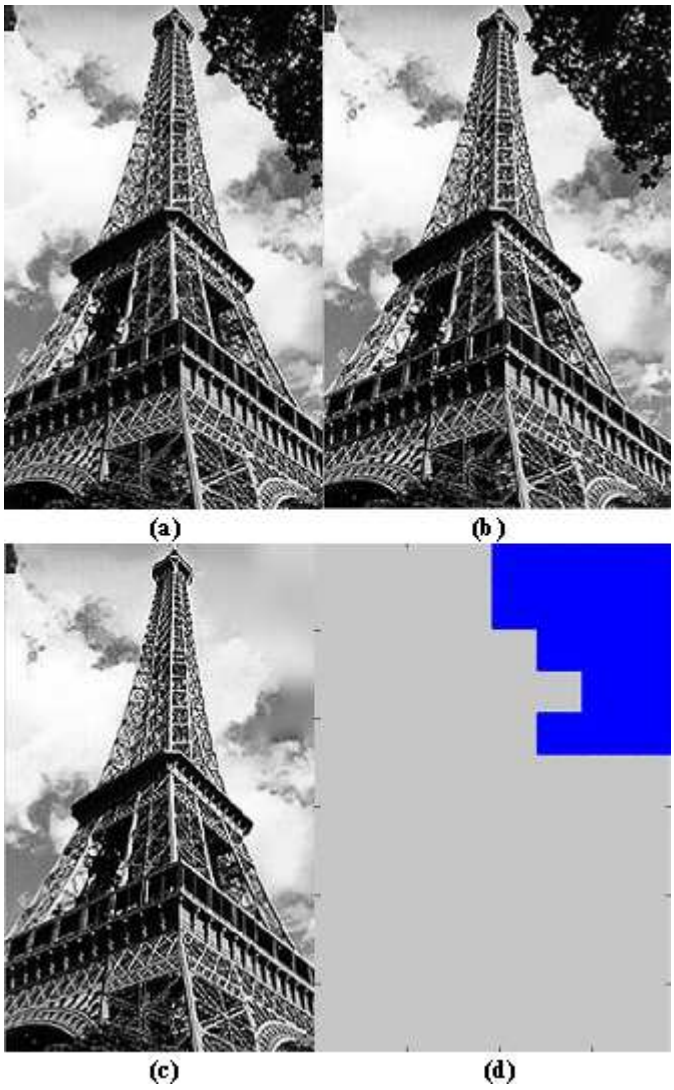


Figure 3: (a) Original Image, (b) Watermarked version, (c) Tampered version: the tree that is shown in upper right corner has been erased, (d) Detection results: the grey area is considered authentic while blue areas are considered tampered.