EURASIP Journal on Applied Signal Processing 2002:6, 613–621

A Survey of Watermarking Algorithms for Image Authentication

Christian Rey
Multimedia Department, Eurecom Institute, 2229 route de Crêtes, B.P. 193, F-06904 Sophia Antipolis, France
Email: christian.rey@eurecom.fr

Jean-Luc Dugelay
Multimedia Department, Eurecom Institute, 2229 route de Crêtes, B.P. 193, F-06904 Sophia Antipolis, France
Email: jean-luc.dugelay@eurecom.fr

Received 20 November 2001 and in revised form 8 March 2002

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. Insuring digital image integrity 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 emerging techniques for detecting whether image tampering has taken place. Compared to the techniques and protocols for security usually employed to perform this task, the majority of the proposed methods based on watermarking, place a particular emphasis on the notion of content authentication rather than strict integrity. In this paper, we introduce the notion of image content authentication and the features required to design an effective authentication scheme. We present some algorithms, and introduce frequently used key techniques.

Keywords and phrases: image processing, security, cryptography, watermarking, content authentication, review, state of the art.

1. INTRODUCTION

1.1. Basic watermarking principles

The digital revolution, the explosion of communication networks, and the increasingly growing passion of the general public for new information technologies lead to exponential growth of multimedia document traffic (image, text, audio, video, etc.). This phenomenon is now so important that insuring protection and control of the exchanged data has become a major issue. Indeed, from their digital nature, multimedia documents can be duplicated, modified, transformed, and diffused very easily. In this context, it is important to develop systems for copyright protection, protection against duplication, and authentication of content. Watermarking seems to be the alternative solution for reinforcing the security of multimedia documents.

The aim of watermarking is to include subliminal information (i.e., imperceptible) in a multimedia document to ensure a security service or simply a labelling application. It would be then possible to recover the embedded message at any time, even if the document was altered by one or more nondestructive attacks, whether malicious or not.

Until now, the majority of publications in the field of watermarking mainly address the copyright of still images. Other security services, such as image content authentication, are still marginal and many fundamental questions remain open. We may wonder, for example, whether it is preferable to use a fragile watermark, a robust watermark, or even use a completely different technique. Furthermore, an authentication service partially calls into question the settings commonly established in watermarking copyright protection, particularly in terms of the quantity and nature of hidden information (for copyright, the mark is independent of the image and is usually a 64-bit identifier), as well as in terms of robustness.

1.2. Notions of integrity

In the security community, an integrity service is unambiguously defined as one, which insures that the sent and received data are identical. This binary definition can also be applicable to images, however it is too strict and not well adapted to this type of digital document. Indeed, in real life situations, images will be transformed. Their pixel values will therefore be modified but not the actual semantic meaning of the image. In other words, the problem of image authentication concerns the image content, for example, when modifications of the document may change its meaning or visually degrade it. 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 such as captions or faces, and manipulations related to the use of an image, such as format conversion, compression, filtering, and so on.

Unfortunately this distinction is not always clear, it partly depends on the type of image and its use. Indeed the integrity criteria of an artistic masterpiece and a medical image will not be the same. In the first case, a JPEG compression will not affect the perception of the image, whereas in the second case it may discard some of the fine details which would render the image totally useless. Even if the scope of this paper is the authentication of multimedia images for general purpose, it is interesting to notice that there exist methods dedicated to very specific integrity services, such as the authentication of medical or military images. Indeed these images should be modified by no means (including watermarking) and a strict definition of integrity is then required. The first class of these methods is invertible watermarking scheme [1], in the sense that, if the image is deemed authentic, the distortion due to the watermarking process can be removed to obtain the original image. Another approach [2] consists in separating the image into two zones: a region of interest (ROI) which is the part of the image used for the diagnostic, where data integrity must be strictly controlled, and a region of noninterest (where distortions are allowed) used to embed the authentication data.

1.3. Classical examples of malicious manipulations

It is a well-known saying that an image is worth a thousand words. Images tend to have more impact on people than text, as it is easier to disregard the content of textual information than to question the origin and authenticity of a *photograph*. It used to be stated that the camera could not lie. However, it is now possible to edit pictures easily and at very little cost. The resulting images can have such a high quality that they appear to be genuine.

In this context, it is obvious that an image authentication service cannot be used to verify the events, but it may be able to detect an a posteriori alteration to an image (i.e., the difference between the photograph as taken, and its released version).

Recently, a picture published on the front page of the Austrian newspaper *Neue Kronen Zeitung*, claims to illustrate that the demonstrators opposed to Haider's party joining the government were aggressive.

Using digital modification, the picture was cropped and the distance between a demonstrator and a policeman was reduced, so that it seemed that the policeman had been struck. In reality, there was approximately two meters between the two persons as certified by the original picture published by the Reuters agency http://www.reuters.com.

The use of image, audio, or video elements in legal situations becomes more and more questionable at a time where surveillance video cameras are increasingly common in towns and other public places.

1.4. Generic image authentication system

Various formulations have been proposed by Wu and Liu [3] and Lin and Chang [4].

However, we propose a generic image authentication system. To be effective, a system must satisfy the following criteria:

(1) Sensitivity: the system must be sensitive to malicious manipulations (e.g., modifying the image meaning) such as cropping or altering the image in specific areas.

(2) Tolerance: the system must tolerate some loss of information (originating from lossy compression algorithms) and more generally nonmalicious manipulations (generated, e.g., by multimedia providers or fair users).

(3) Localisation of altered regions: the system should be able to locate precisely any malicious alteration made to the image and verify other areas as authentic.

(4) 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 was the original content of the manipulated areas.

In addition, some technical features must be taken into account:

(i) Storage: authentication data should be embedded in the image, such as a watermark, rather than in a separate file, as is the case with an external signature.

(ii) Mode of extraction: depending on whether authentication data is dependent or not on the image, a full-blind or a semiblind mode of extraction is required. It is quite obvious that a nonblind mode of extraction does not make sense for an authentication service, since the original image is necessary.

(iii) Asymmetrical algorithm: contrary to classical security services such as copyright protection, an authentication service requires an asymmetrical watermarking (or encryption) algorithm (i.e., only the author of an image can secure it, but any user must be able to check the content of an image).

(iv) Visibility: authentication data should be invisible under normal observation. It is a question of making sure that the visual impact of watermarking is as weak as possible so that the watermarked image remains faithful to the original. Recently, a new approach based on invertible algorithms [1] has been proposed. The basic idea is to be able to remove the distortions due to the watermarking process to obtain the original image data. Obviously perfect in terms of visibility, it is important to note that such an approach could create a very attractive context for attackers.

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

(vi) Protocols: protocols are an important aspect of any image authentication system, in particular avoid protecting a corrupted picture. It is obvious that any algorithm alone can not guarantee the security of the system. It is necessary to define a set of scenari and specifications describing the operation and rules of the system, such as the management of the keys or the communication protocols between owner, seller, client, and so forth.

2. STATE OF THE ART

2.1. Introduction

In this section we do not aim to draw up a complete and exhaustive overview of all image authentication methods. We therefore decided to exclude from this paper any approach which does not include a watermarking aspect, in particular, approaches based on external signature, such as classical cryptographically secure hash functions like MD-4, MD-5 (message digest), CRC-32 (32-bit cyclic redundancy check), SHA-1 (secure hash algorithm) [5], and so on. Interested readers are invited to refer to [6, 7, 8, 9, 10].

Nevertheless, we present a general outline of emerging techniques in order to introduce the key concepts associated with this type of service.

Image authentication systems can be classified in several ways according to whether they ensure strict integrity or content authentication, and also according to the storage mode of data authentication (i.e., watermark or external signature). In this paper, we classify the watermarking methods into two categories (fragile watermarks and semifragile watermarks), even if the concept of robustness is sometimes ambiguous.

2.2. Fragile watermarks

2.2.1 Principle

Most methods currently proposed for providing image authentication are based on a fragile watermark in opposition to robust watermark classically used for copyright protection. The basic idea underlying these techniques is to insert a specific watermark (generally independent of the image data [11]) so that any attempt to alter the content of an image will also alter the watermark itself (Figure 1). 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 these approaches is that it is difficult to distinguish between malicious and nonmalicious attacks (e.g., most fragile methods consider a lossy compressed image as a tampered image, whereas the semantic of the image is unchanged).

2.2.2 Embedding check-sums in LSB

One of the first techniques used for image tampering detection was based on inserting check-sums into the least significant bits (LSB) of the image data. The algorithm proposed by Walton [12] in 1995 consists in selecting, according to a secret key, pseudorandom groups of pixels. The check-sum value is obtained by summing the numbers determined by the 7 most significant bits (MSB) of selected pixels. Then the check-sum bits are embedded in the LSB. The basic version of this algorithm can be summarized as follows.

Algorithm 1 (embedding process).

(1) *Let N be a large integer;*
(2) *divide the image into* 8×8 *blocks;*
(3) *for each block B:*

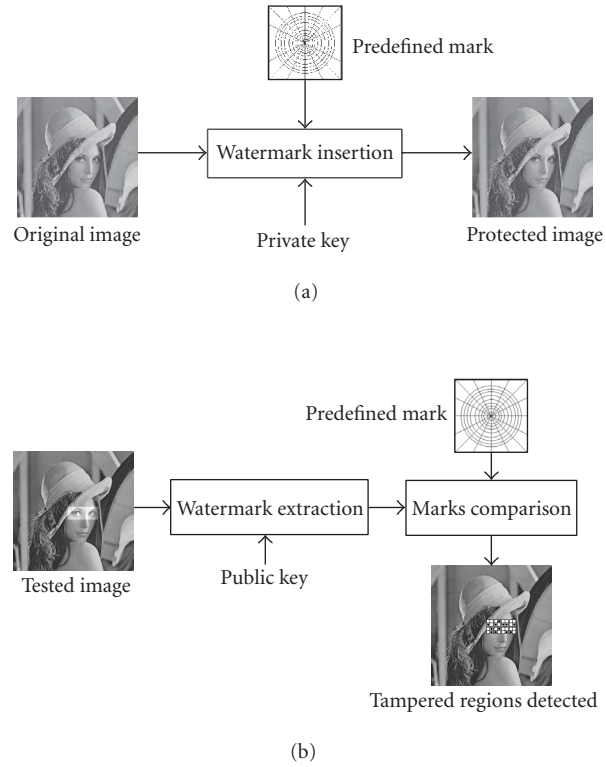

Figure 1: Generic fragile watermark scheme: (a) Image security. (b) Authenticity verification.

(i) *define a pseudorandom walk through all* 64 *pixels, according to the secret key and the block number, and denote the pixels as* $(p_1, p_2, \ldots, p_{64})$;
(ii) *generate a pseudorandom sequence of* 64 *integers* $(a_1, a_2, \ldots, a_{64})$ *comparable in size to N;*
(iii) *the check-sum value S is calculated as*

$$S = \sum_{j=1}^{64} a_j \cdot g(p_j) \bmod N, \tag{1}$$

where $g(p_j)$ is the grey-level of the pixel p_j (determined by the 7 MSB);
(iv) *encrypt the binary form of S;*
(v) *embed the encrypted sequence into the LSB of the image block.*

The checking process is similar to the embedding process. It consists in comparing, for each block, the check-sum determined by the MSB of the tested image with the original check-sum value recovered from the LSB.

The main advantages of this method are that it does not produce visible changes in the image and provides a very high probability of tamper detection. For example, if we swap only two pixels of any block, the check-sum will be modified because each pixel p_j of the block is multiplied by a different coefficient a_j. Furthermore the random walk of the pixels p_j and the coefficients a_j are block dependent, thus making it

impossible to swap or duplicate entire blocks without making undetected changes. One of the drawbacks of this technique is that it is possible to swap homologous blocks (that are blocks of the same position) from two authenticated images protected with the same key. A simple solution to this type of attack is to make the watermark dependent on the image content. This could be achieved using the robust bit extraction algorithm proposed by Fridrich [13].

2.2.3 Self-embedding

Fridrich and Goljan [14] propose an original method for self-embedding an image into itself as a mean of protecting the image content. This method also allows the regions of the image that have been tampered with, cropped, or replaced, to be partially repaired. The basic principle of this method is to embed a compressed version of the image into the LSB of its pixels. As in all watermarking methods based on LSB embedding of the watermark, this method does not introduce visible artefacts. The algorithm consists in dividing the image into 8×8 blocks. Setting the LSB of each pixel to zero and then calculating a DCT (discrete cosine transform) for each block. The DCT matrix is quantified with the quantization matrix corresponding to a 50% JPEG quality. The result is encoded using only 64 bits and the code is inserted into the LSB of another block. The watermarked block must be sufficiently distant from the protected block to prevent simultaneous deterioration of the image and the recovery data during local image tampering. The quality of the recovered regions of the image is somewhat worse than a 50% JPEG quality, but sufficient to inform the user of the original content of these areas. The same authors propose an alternative method, which enables the quality of the reconstructed image to be slightly improved. In this variant, two LSBs are used for embedding the encoded quantified DCT coefficients (i.e., 128 bits can be used instead of 64 bits). For most blocks, 128 bits are enough to encode almost all quantified DCT coefficients. In this way, the quality of the recovered regions is roughly equivalent to a 50% JPEG compression, but due to the modification of the two LSBs, the watermarked image quality is worse.

The major drawback of this method is that the embedded information is not robust. If several distinct regions of the image have been tampered with, the recovery data may also be corrupted. Indeed, after global modifications of the image such as filtering or lossy compression, most reconstruction data will be erroneous as LSB values are changed by this kind of operation.

2.3. Semifragile watermarks

A semifragile watermark is another type of authentication watermark. Semifragile watermarks are more robust than fragile watermarks and less sensitive to classical user modifications such as JPEG compression. The aim of these methods is to discriminate between malicious manipulations, such as the addition or removal of a significant element of the image, and global operations preserving the semantic content of the image.

The use of such methods is mainly justified by the fact that images are generally transmitted and stored in a compressed form. Moreover, for the majority of the applications, the losses due to the compression process do not affect the integrity of the image within the meaning of its interpretation.

2.3.1 Semifragile methods robust to JPEG compression

Lin and Chang [4] propose a semifragile watermarking algorithm that accepts JPEG lossy compression and rejects malicious attacks. They have highlighted and shown two invariance properties of DCT coefficients with respect to JPEG compression.

The first property shows that if we modify a DCT coefficient to an integral multiple of a quantization step Q'_m, which is larger than the steps used in later JPEG compressions, then this coefficient can be exactly reconstructed after JPEG compression.

The second is an invariant relationship between two homologous coefficients in a block pair before and after JPEG compression. Because all DCT coefficients matrices are divided by the same quantization table in the JPEG compression process, the relationship between two DCT coefficients of the same coordinate position from two blocks will not be changed after the quantization process. The only exception is that strict inequalities may become simple equalities due to quantization.

The authentication system proposed by Lin and Chang is based on those two properties. The first one is used to embed the signature and the other is used to generate the authentication bits. The steps of embedding and authentication can be summarized as follows.

Algorithm 2a (generation of authentication bits).

(1) *Divide the original image into* 8×8 *blocks;*
(2) *form block pairs using a predetermined secret mapping function;*
(3) *for each block pair* (p, q)*:*
  (i) *select a set* B_p *of n DCT coefficients;*
  (ii) *generate the binary signature* ϕ_p *of the block pair such that*

$$\phi_p(v) = \begin{cases} 1, & F_p(v) - F_q(v) \geq 0, \\ 0, & F_p(v) - F_q(v) < 0, \end{cases} \tag{2}$$

  where $v \in B_p$, $F(v)$ *is the value of* v*;*
  (iii) *embed the authentication bits according to Algorithm 2b (embedding process).*

The binary signature is then partly embedded into each of the two blocks of the pair. For instance, if the signature length is 8 bits, each block has to embed 4 authentication bits. The embedding process is relatively simple. It consists in defining an equality relation between the LSB of preset DCT coefficients and the bits of the signature.

Algorithm 2b (embedding process).

(1) *Select a set E_p, of n/2 DCT coefficients, where $E_p \cup B_p = \emptyset$;*
(2) *to hide an authentication bit $\phi_p(v)$ into a DCT coefficient ω let*

$$f'_p(\omega) = \left[\frac{F_p(\omega)}{Q'_m(\omega)}\right],$$

$$\tilde{F}_p(\omega) = \begin{cases} f'_p(\omega)\cdot Q'_m(\omega), & \text{if LSB}\left(f'_p(\omega)\right) = \phi_p(v), \\ \left(f'_p(\omega) = \text{sign}\left(\dfrac{F_p(\omega)}{Q'_m(\omega)} - f'_p(\omega)\right)\right)\cdot Q'_m(\omega), & \text{otherwise,} \end{cases} \tag{3}$$

where $\text{sign}(x) = 0$ *if* $x < 0$, 1 *otherwise.*

The authentication process consists in first extracting the authentication bits from the watermarked areas of the image and using them to verify whether the DCT coefficient relationships in the signature match the predicted criteria. If they match, the image is considered authentic. If they do not, this means that either block, or possibly the two blocks, of the considered pair has been manipulated.

The authors have proposed some improvement such as recovery bits. The advantage of these overhead bits is twofold. On the one hand, they allow an approximation of the original block to be reconstructed, on the other hand they help to locate precisely the zones of the images which were really faded (i.e., to raise the ambiguity of the identification of the altered blocks). The recovery bits are generated from a down-sampled and compressed version of the original image. They are then embedded into 4 blocks. The embedding process of recovery bits is similar to that of authentication bits.

2.3.2 Block-based watermark

Block-based watermarking techniques consist in dividing the image into blocks of about 64 × 64 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, we can affirm that the tested image is authentic.

The variable-watermark two-dimensional technique (VW2D) described by Wolfgang and Delp [9, 10] is based on the principle described previously. A binary watermark $W(b)$ is embedded in each block b of an image X. Like Van Schyndel et al. [15], the authors recommend to use m-sequences [16] 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:

$$Y(b) = X(b) + W(b), \tag{4}$$

where X is the original image, and Y the watermarked image.

The verification process used to test if an image Z is authentic consists in computing a statistic score d (6) based on a spatial cross-correlation function:

$$R_{AB}(b) = \sum_{i=0}^{b_{\text{width}}} \sum_{j=0}^{b_{\text{height}}} A(i,j)B(i,j), \tag{5}$$

$$\delta(b) = R_{YW}(b) - R_{ZW}(b), \tag{6}$$

where Z is the tested image (the watermark W is supposed to be known).

If $d < T$, where T is a user-defined threshold, the tested block is considered genuine. While modifying the value of T, one tolerates more or less significant changes in the image. It is then possible to refine detection by defining several thresholds corresponding to several levels of block degradation (e.g., unaltered, slightly altered, very altered, completely changed).

However, in practice, this method offers only a limited interest insofar as it is necessary to store at least, for each block b of an image, the result of the correlation between the watermarked block $Y(b)$ and the watermark $W(b)$.

Fridrich [17, 18] proposes a similar technique. To prevent unauthorized removal or intentional watermark distortion, the author recommends to make the mark dependent on the image in which it is embedded. The binary mark used corresponds to a pseudo-random signal generated from a secret key, the block number and the content of the block represented with an M-tuplet of bits. Each block is then watermarked using Ó Ruanaidh spread spectrum technique [19]. The author claims that the watermark is fairly robust with respect to brightness and contrast adjustment, noise adding, histogram manipulation, cropping, and moderate JPEG compression (up to 55% quality). These watermark properties enable us to distinguish malicious manipulations from visible nonmalicious changes due to common image processing operations.

2.3.3 Feature-based watermark

The basic idea of this method [20, 21] 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 was not tampered with, otherwise the differences will indicate the altered areas (Figure 2).

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 (paintings, satellite images, medical images, and so on). The features are typically selected so that invariant properties are maintained under weak image alterations (lossy compression) and broken for malicious manipulations. These features could be also used to partially restore the tampered regions of the image. Typical features used to provide image authentication are edges, colours, gradient, luminance, or combinations of these features.

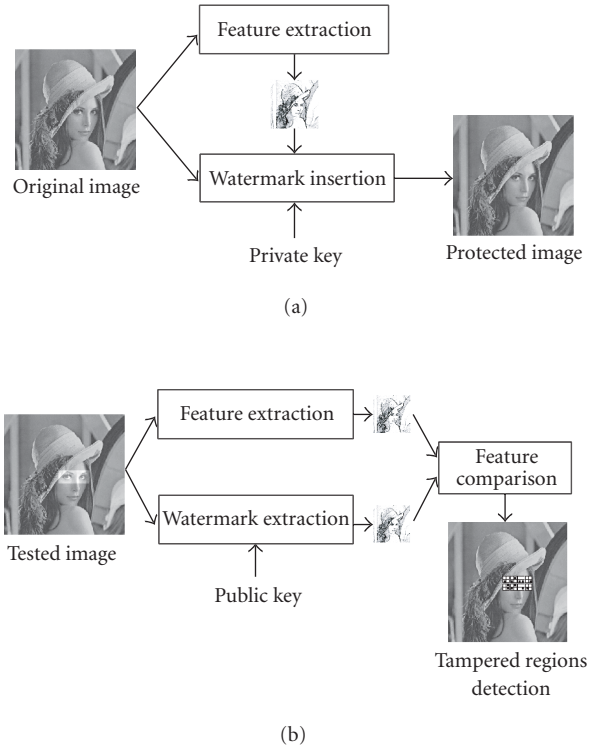

FIGURE 2: Generic semifragile watermark scheme: (a) Image security. (b) Authenticity verification.

A number of constraints are imposed by this method, mainly in terms of robustness and storage capacity of the signature. Robustness is required in order to allow lossless extraction of the watermark. The accuracy of the detection and the amount of information inserted into the image are directly related. It is necessary to find a good compromise for the size of the signature so that both robustness and accurate detection can be achieved.

One of the problems faced by this method is that the image is slightly modified while inserting the watermark. Even slight image variations may affect the image properties. Thus since the features of the original image and the watermarked image are not exactly the same, there are risks of false positive detection. This risk may be more or less important according to the choice of selected features. In order to solve this problem we have implemented an iterative watermarking algorithm. The idea here is to sign the image, extract features from the newly obtained image, and then repeat the watermarking process on the original image (in order to avoid cumulating distortions) using the newly computed features. This iterative process enables hidden features to perfectly coincide with the protected image features. In practice, three iterations are enough.

2.3.4 Other approaches

Other techniques are studied or investigated. Kundur and Hatzinakos [22], and Lin and Chang [23] propose wavelet-based image authentication. The principle of the Lin and Chang method consists in first choosing a wavelet basis and a pseudo-noise pattern (e.g., a 16 × 16 pixels pattern spatially repeated in the horizontal and vertical directions) selected according to a secret key. The image is then decomposed into four sub-bands, LL, LH, HL, and HH, using the previously designated wavelet basis. The HH subband is substituted by the pseudo-noise pattern. Lastly, the watermarked image is obtained after applying the inverse wavelet transformation. Note that the embedding process changes only the HH subband of the image (i.e., high frequencies) and that it does not introduce important visual degradation to the image.

The authentication process is based on the detection of the presence of the embedded pseudo-noise pattern. The first step consists in extracting the HH subband. The extracted subband is then convolved with the pseudo-noise pattern. If the image was not manipulated, the convolution result should be like a dot matrix. In the opposite case, the distribution will lose its uniform character in the areas where the image was tampered with. The authors point out that their method is robust with some filtering operations such as blurring and edge enhancing, and with a soft JPEG compression. On the other hand, the authors do not show the robustness of their method versus specific attacks such as the substitution or preservation of the HH watermarked subband. In other words, is the choice of the wavelet basic as secret, sufficient to avoid this type of attack?

2.4. Summary of different methods

We summarize the different methods presented in this article in Table 1 below. The class to which each method belongs is indicated: fragile, semifragile, digital signature, as well as the type of authentication data used, the authentication data support, the objectives regarding integrity (i.e., strict or content), and whether the method offers a possible localisation and/or reconstruction of the areas tampered with.

By analyzing this table, we can notice that generally the fragile watermarking methods allow only a strict integrity service, whereas the semifragile watermarking methods and the methods based on external signature guarantee a content authentication. However, the fragile watermarking methods remain the simplest to implement.

It is also interesting to notice that only few methods are currently able to restore, even partially, the tampered regions of the image.

3. MALICIOUS ATTACKS

Our aim in this section is not to develop a list of all the possible malicious attacks that an image authentication system can overcome, but to show some of the most frequent attacks. The common objective of these attacks is to trick the authentication system, in other words, to show that an image as authentic even though its content has been modified (or sometimes, the opposite). Some of these attacks look trivial and easy to avoid; nevertheless, it is very important to take them into account when developing an authentication algorithm.

Table 1: Summary of methods ensuring an authentication service.

Method	Class	Mark[1]		Cover	Integrity[2]	Localisation	Reconstruction
Yeung and Mintzer [11]	fragile	predefined logo	no	pixels	strict	yes	no
Walton [12]	fragile	checksums	yes	LSB	strict	yes	no
Fridrich and Goljan [14]	fragile	image comp.	yes	LSB	strict	yes	yes
Wong [24]	fragile	hash function	yes	LSB	strict	yes	no
Lin and Chang [4]	semifragile	DCT coef.	yes	DCT	content	yes	yes
Wolfgang and Delp [9] (1)	semifragile	m-sequences	no	pixels	content	yes	no
Rey and Dugelay [21]	semifragile	luminance	yes	IFS	content	yes	yes
Fridrich [17, 18]	semifragile	block-based	yes	pixels	content	yes	no
Kundur and Hatzinakos [22]	semifragile	random noise	no	wavelets	strict	yes	no
Lin and Chang [23]	semifragile	random noise	no	wavelets	content	yes	no
Queluz [25]	signature	edges	yes	external	content	yes	no
Bhattacharjee and Kutter [6]	signature	interest points	yes	external	content	yes	no
Lin and Chang [7, 8]	signature	DCT coef.	yes	external	content	yes*	no
Wolfgang and Delp [9] (2)	signature	hash function	yes	external	strict	yes*	no

[1] indication whether authentication data is dependent on the image or not.
[2] indicating sensitivity to JPEG compression.
* ambiguity in locating areas that have been tampered with.

One of the most common attacks against fragile watermarking systems consists of trying to modify the protected image without altering the embedded watermark, or even more common, trying to create a new watermark that the authenticator will consider as authentic. Take the following simplified example: the integrity of an image is insured by a fragile watermark, independent of the image content, embedded in the LSB of its pixels. We easily see that if we modify the image without taking account of which bits are affected by this manipulation, we will most likely degrade the watermark and therefore the attack will be detected. On the other hand, if we alter the image without modifying the LSB; the watermark will remain as it was, and the authentication process will not detect any falsification.

In general, when the integrity of an image is based on by a mark that is independent of its content, it is possible to develop an attack that could copy a valid watermark of one image into another image. By doing so, the second image becomes protected even though the second image is false. This attack can even be performed over the same image. First, extract the watermark from the image; then manipulate the image, and finally reinsert the watermark on the altered image. This process will cheat the authentication system.

Following the same philosophy, the *Collage-Attack* proposed by Fridrich et al. [26] creates a falsified image from parts of a group of images protected by the same authenticator using the same mark and the same key. This attack does not assume a priori any knowledge about the hidden binary watermark, or the secret key. Its principle is relatively easy since it replaces each pixel of the altered image by the closest pixel value of equal coordinates of the images in the base. The main difficulty of this method lies on obtaining a database of images rich enough to obtain a falsified image of good visual quality.

Another classic attack tries to discover the secret key used to generate the watermark. This kind of attack, also called *Brute Force Attack*, is very well known by the security community. Once the key has been found, it is very easy for a "hacker" to falsify a watermark of an image that has been protected by this key. The only way to counter this attack is to use long keys to dissuade the attacker from trying to discover the key, because of the high cost of computing time.

Lastly, it is interesting to notice that protocol attacks are also investigated. In [27] Radhakrishnan and Memon propose an attack against the image authentication system SARI [28]. The authors show that the image digest of the SARI system is not secure under certain circumstances. Specifically, if an attacker has the image digests for a multiple number of images where the same secret key has been used to generate the digest, he is able to cause arbitrary images to be authenticated. The authors propose several countermeasures to overcome this attack.

4. 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. There is such a strong demand for image manipulation techniques and applications that they are becoming more and more sophisticated and are accessible to a greater number of people. An unfortunate consequence of this is that new specialized counterfeiters have appeared. Image watermarking, although being a very recent field of research, can propose complementary counterattack methods to the classical cryptographic ones. Its approach grants priority to

the content authentication more than to the strict digital integrity.

In the current state of research, it is difficult to affirm which approach seems most suitable to ensure an integrity service adapted to images and in a more general way to multimedia documents. There does not exist, for the moment, any solution perfectly answering this problem. Fragile watermarking methods are very sensitive to the slightest deterioration of the image, but they offer only a strict integrity service, relatively far from users' needs. Nevertheless, the advantage of fragile watermarking techniques, compared to the methods classically used in security, is that they allow a precise localisation of the manipulated areas. However, the current tendency is more and more towards the use of semifragile methods. These methods are much more tolerant in respect of nonmalicious manipulation, such as a good quality JPEG compression. This flexibility is made possible partly due to watermarking algorithms designed with specific robustness criteria (i.e., the mark is resistant only to certain well-defined manipulations), and also to the use of invariant authentication data to modification preserving the semantic content of the image. The use of a mark dependent on the image content allows, on the one hand, to increase the robustness of the method in respect of malicious attacks, such as the *Collage-Attack*, and on the other hand, a possible partial repair of the altered areas, according to the chosen features.

Generally speaking watermarking research lacks of a rigorous theoretical framework until now. But, follow some empirical results already available, very recent works dealing with theoretical aspects of watermarking appear within the community. In [29], Martinian et al. present one information theoretic formulation of the multimedia authentication problem. They highlight a link between multimedia authentication and a wide array of powerful results from signal processing and information theory. They examine in particular the use of error-correcting codes in authentication.

Additionally, digital signature methods offer an interesting alternative to classical watermarking techniques, insofar there is no longer a limitation in terms of capacity, nor a problem of robustness, thus offering better localisation of the manipulated areas, better quality reconstruction, and a limited risk of false alarms. Moreover, there is already a high level of expertise in the area of community security. However, the major drawback of these techniques is that the image alone is not self-sufficient. Therefore, the benefits of watermarking are reduced and it becomes necessary to be able to guarantee the authenticity of the image/signature pair. Moreover digital signature methods are not very practical to use with multimedia documents. Finally, future developments should not exclude methods based on the combination of robust watermarking and external signature methods. Watermarking would just be an identifier which would allow a trusted user access to the registered signature [30].

Before concluding, it is interesting to point out that even though current methods designed for image integrity may not be perfect, technical demonstrations [28, 31] and commercial products, software and technical material are already available to the public. The most recent and complete R&D (Research and Development) demonstration is without any doubt SARI (self-authentication and recovery images) which is based on a semifragile watermarking technique [4]. SARI is able to detect malicious manipulations, such as crop-and-replacement, and approximatly recover the original content in the altered areas. Another important feature of SARI is its compatibility to JPEG lossy compression within an acceptable quality range. The main commercial products are: the DSS system from Kodak [32] (Digital Signature Standard, standard recognized by the *National Institute of Standards and Technology*, http://www.itl.nist.gov/fipspubs/by-num.htm), the IAS system (Image Authentication System) from Epson http://www.epson.co.uk/, Veridata from Signum Technologies http://www.signumtech.com/, Eikonamark from AlphaTec Ltd http://www.alphatecltd.com, Mediasign from MediaSec http://www.mediasec.com, and PhotoCheck from AlpVision http://www.alpvision.com. Kodak and Epson systems are directly integrated into their digital cameras in order to protect images as they are digitized. The applications covered by these products are multiple. They range from image authentication for expert needs, to the protection of digital documents, for example, images from security video cameras, in the event that they may be used in court. AlpVision and Signum Technologies propose more original uses such as reinforcing the security of paper documents, for example, passports or badges by watermarking their ID pictures.

REFERENCES

[1] J. Fridrich, M. Goljan, and R. Du, "Invertible authentication," in *Proc. SPIE Conf. Security and Watermarking of Multimedia Contents III*, vol. 4314, pp. 197–208, San Jose, Calif, USA, January 2001.

[2] G. Coatrieux, B. Sankur, and H. Maître, "Strict integrity control of biomedical images," in *Security and Watermarking of Multimedia Contents III*, vol. 4314 of *SPIE Proceedings*, San Jose, Calif, USA, January 2001.

[3] M. Wu and B. Liu, "Watermarking for image authentication," in *Proc. IEEE International Conference on Image Processing*, vol. 2, pp. 437–441, Chicago, Ill, USA, October 1998.

[4] C.-Y. Lin and S.-F. Chang, "Semi-fragile watermarking for authenticating JPEG visual content," in *Proc. SPIE International Conf. on Security and Watermarking of Multimedia Contents II*, vol. 3971, San Jose, Calif, USA, January 2000.

[5] SHA-1, "Secure hash standard (SHS)," specification (FIPS 180-1), April 1995, http://www.itl.nist.gov/fipspubs/fip180-1.htm.

[6] S. Bhattacharjee and M. Kutter, "Compression tolerant image authentication," in *Proc. 5th IEEE International Conference on Image Processing (ICIP '98)*, pp. 435–439, Chicago, Ill, USA, October 1998.

[7] C.-Y. Lin and S.-F. Chang, "Generating robust digital signature for image/video authentication," in *Proc. Multimedia and Security Workshop at ACM Multimedia '98*, Bristol, UK, September 1998.

[8] C.-Y. Lin and S.-F. Chang, "A robust image authentication method surviving JPEG lossy compression," in *Proc. SPIE Storage and Retrieval of Image/Video Database*, vol. 3312, pp. 296–307, San Jose, Calif, USA, January 1998.

[9] 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.
[10] R. B. Wolfgang and E. J. Delp, "Fragile watermarking using the VW2D watermark," in *Security and Watermarking of Multimedia Contents*, vol. 3657 of *SPIE Proceedings*, pp. 40–51, San Jose, Calif, USA, January 1999.
[11] 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.
[12] S. Walton, "Information authentification for a slippery new age," *Dr. Dobbs Journal*, vol. 20, no. 4, pp. 18–26, 1995.
[13] J. Fridrich, "Robust bit extraction form images," in *Proc. IEEE International Conference on Multimedia Computing and Systems*, vol. 2, pp. 536–540, Florence, Italy, June 1999.
[14] J. Fridrich and M. Goljan, "Protection of digital images using self embedding," in *Symposium on Content Security and Data Hiding in Digital Media*, New Jersey Institute of Technology, Newark, NJ, USA, May 1999.
[15] 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.
[16] J. G. Proakis, *Digital Communications*, McGraw-Hill, New York, NY, USA, 3rd edition, 1995.
[17] J. Fridrich, "Image watermarking for tamper detection," in *Proc. IEEE International Conference on Image Processing*, vol. 2, pp. 404–408, Chicago, Ill, USA, October 1998.
[18] J. Fridrich, "Methods for detecting changes in digital images," in *Proc. IEEE International Conference on Image Processing*, Chicago, Ill, USA, October 1998.
[19] J. J. K. Ó Ruanaidh and T. Pun, "Rotation, scale and translation invariant digital image watermarking," in *Proc. IEEE International Conference on Image Processing*, vol. 1, pp. 536–539, Santa Barbara, Calif, USA, October 1997.
[20] J.-L. Dugelay and S. Roche, "Process for marking a multimedia document, such an image, by generating a mark," Pending patent EP 99480075.3, EURECOM 11/12 EP, July 1999.
[21] 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.
[22] D. Kundur and D. Hatzinakos, "Towards a telltale watermarking technique for Tamper-Proofing," in *Proc. IEEE International Conference on Image Processing*, vol. 2, pp. 409–413, Chicago, Ill, USA, October 1998.
[23] C.-Y. Lin and S.-F. Chang, "A watermark-based robust image authentication method using wavelets," ADVENT project report, Columbia University, April 1998.
[24] P. Wong, "A watermark 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.
[25] M. P. Queluz, "Towards robust, content based techniques for image authentication," in *Proc. IEEE Signal Processing Society 1998 Workshop on Multimedia Signal Processing*, pp. 297–302, Los Angeles, Calif, USA, December 1998.
[26] J. Fridrich, M. Goljan, and N. Memon, "Further attacks on Yeung-Mintzer fragile watermarking scheme," in *Security and Watermarking of Multimedia Contents II*, vol. 3971 of *SPIE Proceedings*, San Jose, Calif, USA, January 2000.
[27] R. Radhakrishnan and N. Memon, "On the security of the SARI image authentication system," in *Proc. IEEE International Conference on Image Processing (ICIP 01)*, Thessaloniki, Greece, October 2001.
[28] C.-Y. Lin and S.-F. Chang, "SARI: Self-authentication-and-recovery image watermarking system," in *Proc. 9th ACM International Conference on Multimedia*, Ottawa, Canada, 30 September–5 October 2001.
[29] E. Martinian, B. Chen, and G. W. Wornell, "Information theoretic approach to the authentication of multimedia," in *Proc. SPIE Conf. Security and Watermarking of Multimedia Contents III*, vol. 4314, pp. 185–196, San Jose, Calif, USA, January 2001.
[30] "RNRT-AQUAMARS," National French project (1999–2001), http://www.telecom.gouv.fr/rnrt/wprojets.htm.
[31] C. Rey and J.-L. Dugelay, "Image watermarking for owner and content authentication," in *Proc. ACM Multimedia Technical Demonstration*, Los Angeles, Calif, USA, November 2000.
[32] Kodak, "Understanding and integrating KODAK picture authentication cameras," http://www.kodak.com/US/en/digital/software/imageAuthentication/ White paper.

Christian Rey graduated with master degree in computer science and artificial intelligence from the University of Sciences of Luminy (Marseille), France in 1998. He joined Eurécom in May 1999 as Ph.D. student, registered at the University of Avignon. His research interests currently include image/video watermarking, with a particular emphasis on malicious attacks on watermarking systems and counter-measures, image authentication, and turbo codes. He was involved in the French national research project RNRT AQUMARS, on work package "image authentication." He is the coauthor of several publications, demo, and patents related to watermarking.

Jean-Luc Dugelay is professor in the Department of Multimedia Communications at the Institut Eurécom in Sophia Antipolis, he works in the area of multimedia signal processing. He got his Ph.D. degree from the University of Rennes in 1992. His main research interests are imaging for security (watermarking and biometrics) and face cloning for telecommunications. He contributed to the first book on watermarking (information hiding techniques for steganography and digital watermarking, Artech House 1999). He gave several tutorials on digital watermarking (coauthored with F. Petitcolas from Microsoft Research) at major conferences (ACM MultiMedia, October 2000, Los Angeles, and Second IEEE Pacific-Rim Conference on Multimedia, October 2001, Beijing). He is currently serving as a Consultant in watermarking for France Telecom R&D and STMicroelectronics. His group is involved in the European IST project Certimark (Certification of watermarking techniques). He is Associate Editor of the IEEE Transactions in Image Processing, the EURASIP Journal on Applied Signal Processing and the Kluwer Multimedia Tools and Applications. He is also a member of the IEEE Signal Processing Society, Multimedia Signal Processing Technical Committee (IEEE MMSP TC), and Image and Multidimensional Digital Signal Processing (IMDSP TC). He was technical cochair and organizer of the fourth IEEE workshop on Multimedia Signal Processing, Cannes, October 2001.