Security of Fragile Authentication Watermarks with Localization

Jessica Fridrich*
SUNY Binghamton; Dept. of Electrical Engineering, Binghamton, NY 13902-6000

ABSTRACT

In this paper, we study the security of fragile image authentication watermarks that can localize tampered areas. We start by comparing the goals, capabilities, and advantages of image authentication based on watermarking and cryptography. Then we point out some common security problems of current fragile authentication watermarks with localization and classify attacks on authentication watermarks into five categories. By investigating the attacks and vulnerabilities of current schemes, we propose a variation of the Wong scheme[18] that is fast, simple, cryptographically secure, and resistant to all known attacks, including the Holliman-Memon attack[9]. In the new scheme, a special symmetry structure in the logo is used to authenticate the block content, while the logo itself carries information about the block origin (block index, the image index or time stamp, author ID, etc.). Because the authentication of the content and its origin are separated, it is possible to easily identify swapped blocks between images and accurately detect cropped areas, while being able to accurately localize tampered pixels.

Keywords: Authentication, fragile watermarks, attacks, security, Holliman-Memon attack, localization

1. INTRODUCTION

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

In the past few years, many new techniques and concepts based on data hiding or steganography have been introduced as a means for tamper detection in digital images and for image authentication. Fragile watermarks[1,2,12,15,18,21] are designed to detect every possible change in pixel values. In many schemes, it can be shown that without the secret key, the probability of a modification that will not be detected can be related to a cryptographic element present in the scheme, such as a hash function. Semi-fragile watermarks[10,11,17] are moderately robust and thus provide a "softer" evaluation criterion (authentication with a "degree"). Some schemes have been specifically designed to be compatible with certain distortion, such as JPEG[11] or wavelet[20] compression. There is another special group of authentication techniques that can be termed "content authentication"[11,12,14]. In those schemes, robustly extracted image features are embedded in the image in a semi-robust manner to help identify gross changes in the image.

In this paper, we focus on fragile watermarking authentication techniques that have good localization properties and are cryptographically secure in the sense that there is no information leakage from the scheme and all modifications are detected with very high probability derived from a secure cryptographic element. We are especially interested in techniques that resist as wide a spectrum of attacks as possible and can be used for a wide range of practical applications. In Section 2, we compare authentication using watermarking and encryption pointing out the differences, advantages and disadvantages of both paradigms. In Section 3, we classify attacks on authentication schemes based on what information is available to the attacker. In view of those attacks, the strengths and weaknesses of current fragile authentication watermarks are discussed in Section 4. The techniques are evaluated based on their ability to localize changes, detect modifications, and their resistance to attacks. In Section 5, we propose a new, simple, and fast fragile authentication method that has very good localization properties and is cryptographically secure. The new scheme is a modification of the scheme proposed by Wong[18], in which the content authentication is separated from authentication of the content origin. It is resistant to known attacks and can be used for a wide range of applications, such as image authentication in digital surveillance cameras. The paper is concluded in Section 6.

* fridrich@binghamton.edu; phone: 1 607 777-2577; fax: 1 607 777-2577; http://www.ssie.binghamton.edu/fridrich; SUNY Binghamton; T. J. Watson School of Engineering, Dept. of Electrical Engineering, Binghamton, NY USA 13902-6000

2. IMAGE AUTHENTICATION USING WATERMARKS AND CRYPTOGRAPHY–COMPARISON

Message authentication using cryptography starts with computing a message digest using a hash function $H=H(M)$. The hash H is further encrypted with the private (secret) key S of the sender (author or sensor) and appended to the message to obtain the bit-stream $M\&E_S(H)$. In the expressions above, M stands for the message, & denotes the concatenation operation, and E_S is the encryption function with key S. For a secure exchange of data between subjects, this bit-stream can be encrypted with the public key P of the recipient and the final encrypted bit-stream $E_P[M\&E_S(H)]$ is sent to the recipient. The recipient can verify the message integrity and origin by first decrypting the encrypted message using his secret key, decrypting the hash using the public key of the alleged sender, and matching the extracted hash with the hash of the decrypted message body M. This system also prevents framing another person and pretending that a message came from him because the sender's private key is used to encrypt the hash.

While cryptography protects the communication channel, it provides no further protection or evidence of tampering after the image is decrypted. For example, one may desire that the authentication scheme localize the modifications or even estimate their extent. Also, because images have redundant and/or perceptually irrelevant components, it would be desirable to distinguish between malicious tampering, such as feature removing or adding, from innocuous distortions, such as lossy compression or common image enhancement. Furthermore, we may desire to establish the content authenticity of image fragments obtained by cropping. To achieve these goals, we can attempt to adjust the cryptographic authentication protocol and the calculation of the message authentication code (MAC). For example, one can build-in the localization ability by hashing smaller blocks (or other localized image components, such as wavelet coefficients) rather than the whole image, and append this block-wise MAC to the image. It is also possible to quantize the image (or its components) before hashing to allow some small amount of distortion without violating the image authenticity. One could also extract some salient image features from the image in a "robust" way and hash them to build a MAC. In all cases, the MAC is encrypted and appended to the image either in its header or in a separate file. There are some advantages of having this authentication information appended to the image because the appended data does not change with modifications that occur to the image. Thus, one can readily verify the integrity of the image by calculating the MAC from the distorted image and comparing it with the appended MAC. However, there are some definite disadvantages of the MAC appended to the image that justify alternative approaches, such as watermarking.

Many watermarking applications start with the same approach – calculating the MAC from the image. The MAC can be a sensitive function of the image or its blocks (hash) or it can be a set of robust features, such as the edge map, rescaled version of the same image, wavelet maxima, the visual hash[4], etc. The difference is that the MAC is *embedded* in the image rather than *appended* to it. Care needs to be taken so that the same hashes or features can be extracted from the watermarked image. For fragile schemes, this is usually done by dividing the image into two non-interacting parts and assigning one part (e.g., the 7 most significant bits of pixels) to be hashed, while the other part (e.g., the LSBs) are used to embed the hash. Recently, an alternative approach was developed in which the hash is calculated from the whole image (or a block) and embedded in the image in an “invertible” manner[7]. By invertible, it is meant that the exact copy of the original image can be reconstructed after the hash is extracted from the watermarked image. This approach utilizes the concept of “lossless” data embedding and will not be discussed in this paper.

Most watermarking techniques utilize the fact that images contain some perceptually irrelevant information and replace it with the MAC. Thus, contrary to the approach in which the MAC is appended, the overall amount of data that needs to be stored does not change. Also, because the authentication data is tied directly to the image, lossless format conversion, such as changing the format from PGM to BMP, is not indicated as tampering. MAC that is stored in the header or a separate file can be easily lost during format conversion or simple resaving. Another important advantage of authentication based on watermarking is that, if the authentication information is lumped and localized in the image, one can verify the content integrity of image *fragments* after cropping. The watermarking authentication technique proposed in this paper separates the authentication of image content and its origin thus enabling detection of collages from multiple images formed by cutting and pasting. Without special compliant software, this capability can never be achieved with MACs that are appended to the image. Finally, we remark that the MAC that is embedded in the image is not obvious, which in some situations can be an important factor.

Authentication of images based on watermarking has not been intended and cannot replace classical cryptographic authentication protocols. Watermarking cannot provide equivalent satisfactory properties for the scenario in which subjects exchange data using public-key encryption infrastructure. This is intuitively clear because the authenticated image is usually perceptually equivalent to the original, and hence an attacker can simply overwrite the image with his watermark and the resulting image will appear to be authenticated by the attacker. Although this problem can be alleviated to some extent by embedding in the image the sender's ID in a robust manner, the security this mechanism is

nowhere close to what can be obtained using cryptographic authentication. The authentication watermark can, however, be useful in those cases when the set of valid authentication keys is well defined and controlled, such as keys wired in a digital camera. An example scenario for this application is given in Section 3.

There are also some drawbacks of authentication based on watermarking. The authenticated image will inevitably be distorted by some small amount of noise due to the authentication itself. In almost all watermarking schemes, this distortion cannot be completely removed even when the image is deemed authentic. Although the distortion is often quite small, it may be unacceptable for medical imagery (for legal reasons) or military images inspected under non-standard viewing conditions (extreme zoom or enhancement). Recently, lossless (or erasable) authentication watermarks have been proposed[7]. In those schemes, if the image passes the authenticity check, the authentication watermark can be completely removed and the exact original image data obtained.

Another potential disadvantage of watermark-based authentication is that methods developed for uncompressed image formats cannot be easily adapted for lossy formats and usually different techniques must be developed for each lossy format.

In the next section, we classify the attacks on authentication techniques and formulate a practical application scenario in which all of the proposed attacks are indeed plausible.

3. ATTACKS ON AUTHENTICATION WATERMARKS FOR IMAGES

As in cryptography, one can think of several different levels of attacks based on the information and devices available to the attacker. Because the authentication watermark is fragile, the attacker is not interested in making the authentication watermark unreadable. Actually, disturbing the watermark is quite easy because of its fragility. The goal of the attacker is the opposite when compared to a robust watermark. Below are listed three common security problems of most previously proposed schemes. We further describe five attacks arranged in the increasing order of strength.

Undetected Modifications. The attacker is trying to make a change to the authenticated image that will not be detected by the algorithm. He may even be satisfied with making changes that will not be detected with a "reasonable" probability or changes that will be misinterpreted by the detector, such as masquerading cutting and pasting as tampering at the border of the cropped area. One of the most serious problems of practically all schemes that watermark small independent units (groups or blocks) is the Holliman-Memon attack[8] in which the units are exchanged either within one image or among multiple authenticated images.

Information Leakage. Another potential problem that many authentication watermarks have is information leakage. The attacker may be interested in obtaining some information about the secret authentication key, including the placement of MAC in the image pixels, detecting synchronization patterns, deriving portions of look-up tables, or obtaining some statistical evidence about the secret key or entities derived from it, such as a random walk through the image.

Protocol Weakness. Even a scheme that does not have any information leakage and detects all modifications with very high probability may be vulnerable in certain situations. For example in the situation when an attacker has unlimited access to the verification device and is able to submit images for integrity verification (oracle attack).

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

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

Multiple Stego-Image Attack. The attacker has multiple authenticated images and is interested in making undetected changes or recovering information from the scheme.

Verification Device Attack. The attacker has access to the verification device, i.e., the attacker can verify the authenticity of any image. The strength of this attack depends on the output available to the attacker. The output could be a binary Yes/No for the whole image or it could be a bitmap with pixels/blocks indicated as authentic or tampered. Again, the attacker is interested in making undetected changes or recovering secret information from the scheme.

Cover-Image Attack. The attacker has multiple pairs of original-authenticated images. This assumption is not that unreasonable if an attacker can somehow get access to the raw images before authentication has occurred or when plausible statistical hypothesis can be made about the original. Image semantic can also be used to obtain an estimate of the original image. Again, the attacker is interested in making undetected changes or recovering information from the scheme.

Chosen Cover-Image Attack. The attacker has access to the authentication device and can submit his images for authentication. Can he recover some information about the secret authentication key?

In most previously proposed schemes, the authors typically consider only the first type of attack and most schemes are, indeed, resistant to such an attack. However, when multiple images are authenticated with the same key, some schemes begin to exhibit serious security problems. Moreover, virtually all schemes appear to be vulnerable to some degree to at least one of the attacks described above. One of the goals of this paper is to design a secure fragile authentication watermarking scheme that can localize tampering and is not vulnerable to any of the attacks above. We want the new scheme to be usable in practical situations, such as the one described below.

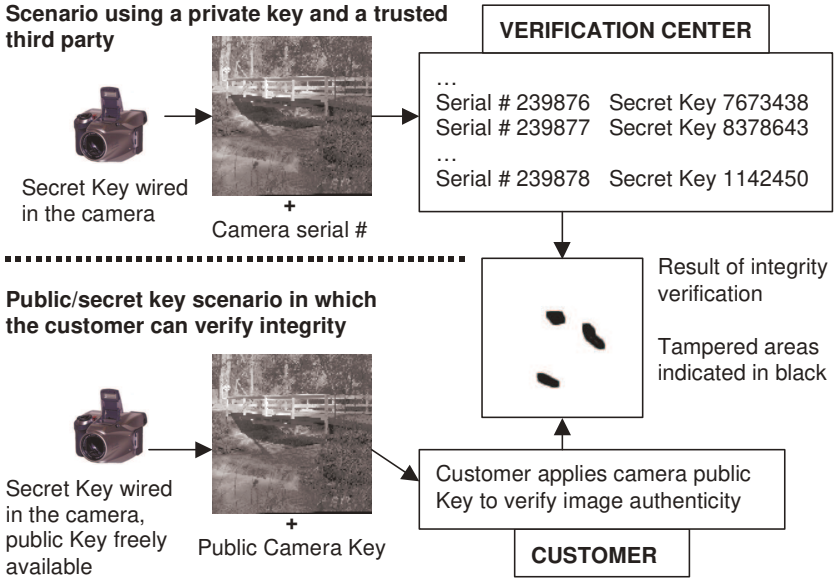

Figure 1 Example of an application scenario in which images are authenticated in the camera with a fragile watermark computed from the secret camera key (and possibly other data provided by the author). The verification center or the customer can perform authenticity check on the image.

In this application scenario, it is assumed that we have possibly many digital cameras equipped with a watermarking chip that authenticates every image the camera takes before storing it on its flash card. The authentication watermark could be uniquely tied to the camera's serial number via a secret key hard-wired in the camera itself thus creating a link between the images and the hardware that took them. The watermark can also depend on some key information provided by the person who took the image. In the future, iris or thumbprint scans as well as simple punch-in passwords could be used to facilitate the connection between a digital image and the person who took them. The fragile watermark could later be used to prove the integrity of a specific image and certify that it has been taken with certain hardware by a specific person. This application scenario is schematically explained in Figure 1. The upper half shows the scenario in which a secret camera key is used in connection with a third trusted party – the verification center – that has the database of cameras' serial numbers and their secret keys. Anybody can submit an image and a camera serial number to check the image integrity and whether or not the image has been taken with that digital camera (and by a specific person, if the watermark was designed to carry that data as well). In the second scenario, the public key encryption is used to remove the need of the trusted party. The private camera key is again used to watermark the image. Anybody can verify the image authenticity by applying the verification algorithm with a public camera key. This scenario is a special case of the application described by Coppersmith et al.[2]

Note that the scenario above calls for a secure authentication watermark that must resist all five attacks above.

4. SECURITY ANALYSIS OF CURRENT FRAGILE AUTHENTICATION WATERMARKS WITH LOCALIZATION

In this section, we analyze the security of current fragile watermarking schemes designed for authentication with localization. The classification of attacks proposed in the previous section is used to evaluate the security and strength of the techniques and their usability for the scenario shown in Figure 1.

4.1. Derivatives of the Yeung-Mintzer Scheme

One of the most popular (and most attacked) fragile watermarking schemes is the Yeung-Mintzer scheme[21]. The watermarking starts with a secret key that is used to generate a binary valued function f: $\{0, 1, \ldots, 255\} \rightarrow \{0,1\}$, that maps each grayscale level g_{ij} in the range from 0 to 255 to either 1 or 0. For color images, three such functions, f_R, f_G, f_B, one for each color channel, are generated. These binary functions are used to encode a binary logo L. The logo should be kept secret and can also be generated from the secret key or it can have graphical meaning. The gray scales g_{ij} are perturbed to satisfy the following expression for each pixel (i,j)

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

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

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

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

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

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

The Yeung-Mintzer scheme is vulnerable to the Multiple Stego-Image attack if the same logo and key are reused for multiple images. For two images I_1 and I_2 with gray levels $g^{(1)}$ and $g^{(2)}$ watermarked with the same key and logo L, we have

$$f_g(g_{ij}^{(1)}) = Lg_{ij} = f_g(g_{ij}^{(2)}) \text{ for all } (i,j) \ .$$

The last expression constitutes $M \times N$ equations for 256 unknowns f_g. Only two images are needed[5,6] on average to recover over 90% of the binary function f_g. Once the binary function is estimated, the logo can be easily derived. Actually, if the logo is a real image rather than a randomized picture, we can use this additional information to recover any missing part of the binary function f_g. Although the situation becomes more complicated for color images, the attack works even better because up to six colors participate in every equation[6].

To improve the security of the Yeung-Mintzer scheme, Fridrich et al[8] proposed to replace the look-up tables with a symmetric encryption scheme E_K with a private key K. The image is scanned by rows as in the original scheme and each pixel g_{ij} is modified to enforce the following relationship

$$L_{ij} = Parity(E_K\{g_{i-u\,j-v} \mid 0 \le u, v \le a-1\}) \ . \tag{3}$$

In the expression above, the pixels $g_{i-u\,j-v}$, $0 \le u, v \le a-1$ form an $a \times a$ square with g_{ij} at its lower right corner and *Parity*( ) is a binary valued function applied to the output of the encryption function E_K. By *Parity* we understand, for example, the excluded OR of the encrypted bit-stream. The value $a \approx 5$ is recommended as a good compromise between security and

localization. Although this modification somewhat worsens the localization properties, the attack on look-up tables is not feasible anymore. The authors also propose to repetitively embed the image index (time stamp) in the image in 32 randomly chosen pixels in each disjoint block of 32×32 pixels by imposing another condition similar to (3) at those index-bearing pixels. The purpose of this measure is to detect cutting and pasting between images.

The proposed modification becomes, however, very computationally expensive because the encryption function E_K must be evaluated at each pixel in the image. Another drawback is that even the modified scheme cannot be used for the scenario shown in Figure 1 because it is vulnerable to the Verification Device Attack if the verification center returns an image with indicated tampered/non-tampered pixels. This is because the attacker can simply take an arbitrary image and start modifying the pixels in a row-by-row manner and observe the output from the verification center till a non-tampered status for that pixel is reached. The attacker then moves to the next pixel till the whole image is scanned (and authenticated). The verification center will only need to be accessed on average $2MN$ times, where M and N are image dimensions. The culprit is the sequential character of the watermarking scheme, and it appears that no modification of this scheme that preserves the sequential character of the watermarking process can remove this weakness. In our search for a secure fragile watermark with good localization, we turned our focus to block-based schemes with secure cryptographic elements.

4.2. Derivatives of the Wong Scheme

One of the first fragile watermarking techniques proposed for detection of image tampering was based on inserting check-sums of gray levels determined from the seven most significant bits into the least significant bits (LSBs) of pseudo-randomly selected pixels[15]. In this section, we describe the variation by Wong[18] because our new method is based on this technique.

Wong divides the image into non-overlapping blocks of $W \times H$ pixels. The watermarking is done for each block separately. Wong[19] described two versions of this algorithm: private key and public key versions. In the private key version, the seven most significant bits of all pixels in the block are hashed using a secure key-dependent hash. The hash is then XORed with a chosen binary logo and inserted into the LSBs of the same block. Verification proceeds in the reverse order first by calculating the key-dependent hash of the 7 MSBs in each block and XORing them with the LSBs. Comparison with the logo indicates tampered blocks. In the public key version, the 7 MSBs are hashed using a fixed hash, XORed with the logo and then encrypted using a public key encryption method. The encrypted bit-stream is again inserted in the LSBs of the same block. The verification algorithm proceeds by blocks and first calculates the hash of the 7 MSBs of all pixels in that block, XORs the hash with the decrypted LSBs (using the public key) and compares the result with the binary logo.

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

One weakness of this scheme is that it is possible to swap blocks in an image without causing a detectable change if the logo has large areas of black or white pixels[5,9]. This problem could be avoided by including the block position (or block index) to the image data before hashing[19] or using random logos. However, there will still be a problem with swapping identically positioned blocks from several authenticated images. This attack has been for the first time described by Holliman and Memon[9] and further analyzed by Fridrich[5]. If an attacker has a database of images authenticated with the same key, it is possible to take an arbitrary image and modify it to make it authentic. One can divide the image into blocks and for each block perform a search through the blocks in the same position in all database images. The original block is then replaced with the closest match. This vector-quantization type of attack is applicable not only to the Wong's scheme but to many other schemes in which it is possible to identify disjoint groups of image elements that are modified without context. Consequently, most schemes that can localize changes are vulnerable to this attack.

The effectiveness of the attack would be decreased by selecting a larger block. However, this measure will also worsen the localization accuracy. An interesting solution to the Holliman-Memon attack was proposed by Coppersmith et al.[2] The authors proposed overlapping blocks in which the hash is calculated from slightly larger blocks but embedded only in the original smaller blocks. Blocks of 24×24 pixels with overlapping blocks of size 32×32 pixels were proposed. While this scheme can solve the problem with block swapping to some degree, it does so at the expense of losing the localization accuracy. One can easily visualize that modifying pixels that are close to the corners of the 24×24 squares will indicate tamper for four 24×24 blocks together forming a 48×48 square.

We also point out that, in this technique, we may not be able to distinguish cropping from tampering the boundary squares lying on the lines of cutting. Let us take two images authenticated with the same key. We cut an identically positioned rectangular area consisting of 8+3×24+8 pixels by 8+2×24+8 pixels from both images (see the dashed line in Figure 2). The cutting lines follow the boundary of the larger 32×32 squares. The six interior 24×24 squares in the swapped rectangles will be verified as 'authentic' and the verification procedure will only indicate tamper at the lines of cutting (shaded squares). We can easily see that modifying pixels in the middle 16×16 square in each shaded 24×24 square will be detected in the exact same manner as replacing the whole rectangular areas.

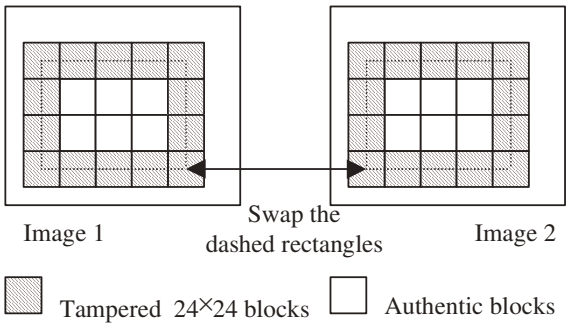

Figure 2 In Wong's scheme with overlapping blocks, swapping the dashed rectangles results in tamper detection in boundary squares (shaded) but the interior 6 squares stay "authentic".

Wong and Memon proposed another approach to solve the Holliman-Memon attack[19]. They recommend including the image index (or time-stamp) in the data before hashing. This measure, indeed, solves the problem and is virtually equivalent to using a different key for each image. In applications, such as when images are distributed via the Internet from a web page, this is actually not a problem because the image index (or its serial number) could be publicly available to assist the customer in verifying the image integrity. In other situations, however, this solution may not be practical because the image index must be known beforehand in order to authenticate the image. For example, in the scenario described in Figure 1, the proposed solution to the attack is not practical because in most situations we will only have the image but have no further information about the image index. All possible indices would have to be tried to see if an authentic image is obtained.

The same authors proposed another idea to thwart the attack by calculating the image index from the hash of all 7 MSBs of all blocks. This approach, however, would lead to a complete loss of localization if the MSBs of at least one block were modified. Including only the MSBs of surrounding blocks would lead to a method similar to the one proposed by Coppersmith et al.[2]

Another possibility would be to embed the image index in a robust manner in the image as the first watermark and overlay the fragile watermark in a second watermarking "pass". This, however, may negatively influence our ability to localize because larger blocks are needed for embedding the index robustly. Depending on the watermarking scheme used, the index embedding will also increase the overall distortion. Finally, adding another element in the scheme will make the authentication algorithm more vulnerable to attacks.

Celik et al.[1] have recently proposed an elegant solution to the Holliman-Memon problem. Their watermark has hierarchical structure, which is essentially a block-wise authentication with highly overlapping blocks. Their solution, however, is more complex than the one proposed in this paper.

The discussion in this section can be summarized as follows. The Holliman-Memon attack still poses a problem for fragile authentication methods whose goal is to accurately localize tampered areas. The method by Coppersmith[2] resists this attack at the expense of losing the accuracy of localization but is not able to distinguish cutting and pasting from pixel tampering at the cutting lines in all cases. In addition to this, the authors believe that a good technique should be able to distinguish cropping from other content modifications. We should be able to recognize if the image has been cropped and how much has been cropped out and, at the same time, be able to authenticate the cropped contents as intact.

In this paper, we propose an elegant and simple technique that successfully addresses the above problems. The idea is based on separating the authentication of the content and the data about the origin of the content. The content

authentication will be achieved using a fixed logo structure, while the logo itself will provide information about the original block position, image index, and other useful data. This idea is elaborated on in the next section.

5. NEW SECURE FRAGILE AUTHENTICATION WATERMARK WITH LOCALIZATION

Based on the security analysis and performance evaluation of current schemes given in the previous section, we propose the following design criteria for the new scheme:

1. Resistance to all five attacks described in Section 3. In particular, we require resistance to the Holliman-Memon attack and the Verification Center Attack, as well as the Cover-Image Attack and the Chosen Cover-Image Attack.
2. Ability to accurately localize modifications with a very high probability derived from a secure cryptographic element.
3. Fast and simple implementation with expandability to audio and other multimedia objects.
4. Ability to distinguish cutting and pasting from cutting-lines tampering and ability to recognize the amount of cropping and original pixel/block positions.

We begin our approach with the Wong scheme and make the following modifications to it. First, we propose to use the binary logo to carry information about the block position and image index and possibly other information relevant to the image, such as its original dimensions, author ID, etc. Furthermore, the block content will be authenticated by creating a simply recognizable *structure* in each logo block rather than the logo content itself. This simple yet powerful idea of separating the content authentication from authentication related to the origin of the content appears to have solved all the problems associated with the Wong scheme, including the problem with the availability of the image index. Also, the new scheme appears to be resistant to all five attacks from Section 3.

Watermark embedding

1. Divide the image into 8×16 blocks (128 pixels)
2. For each block, calculate the hash of the 7 MSBs of all 128 pixels
3. XOR the Hash with the corresponding binary logo block, encrypt, and insert in the LSBs of the same block.

The embedding process is identical to the Wong's technique (see Figure 3). The main difference is the construction of the logo. Its structure is described in Figure 4. Instead of a fixed pattern, the logo is divided into 8×16 blocks, each block carrying information about original block position, image index, original image dimensions (resolution), camera ID, author ID (PIN), random bits, etc. The first 52 bits are the same as the next 52 bits, which creates a structure in the logo that can be used to verify the authenticity of any 8×16 block independently of its position. The block content itself provides information about the block position and origin.

Verification algorithm

1. Divide the image into 8×16 blocks (128 pixels)
2. For each block, calculate the hash H of the 7 MSBs of all 128 pixels
3. Decrypt the LSBs and XOR with H
4. If the first 52 bits are exactly the same as the next 52 bits, the block *content* is authentic (but maybe not its position and context), otherwise it is not (the probability of false authentication is $1{:}2^{52} \approx 10^{-15}$)
5. Authentic blocks give us information about the image index, the original position of the block and the original dimensions of the image. Swapped blocks and cropping can thus be readily detected. Information about the hardware/person that took the picture can also be stored in the logo if larger blocks are used.

If no authentic block is found, one can attempt to use a brute-force search for authentic blocks by sliding an 8×16 window across the image. The size of the image block can be adjusted if authentication with smaller error probability is desired or if more ancillary information needs to be embedded in each block.

We note that without any information about the authentication (encryption) key, the five attacks from Section 3 do not pose any threat to this scheme. In order to create an authentic block, one would have to keep submitting an image to the verification center till a match is obtained. Because a secure cryptographic hash and encryption is used in the scheme, no information leakage occurs and all modifications will be detected with very high probability. Finally, we note that the scheme can be formulated with a secret key (if a private key symmetric encryption function E is used) or a public/secret key if a public key encryption E is used.

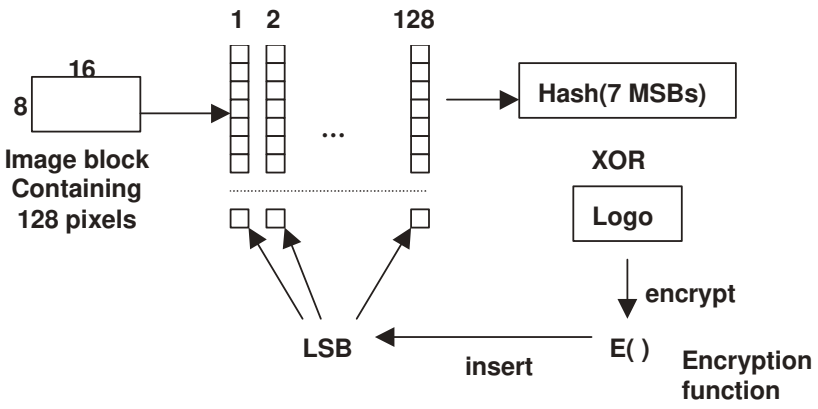

Figure 3 Block diagram for watermarking (the proposed method).

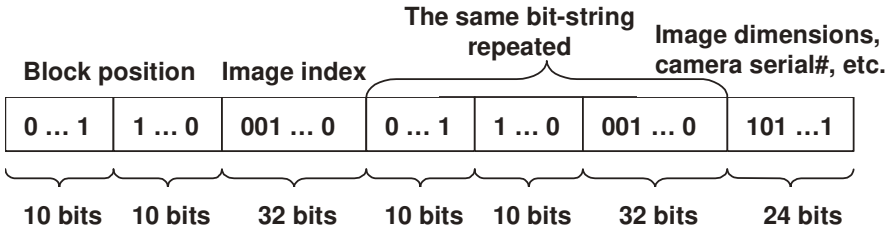

Figure 4 Proposed structure of the binary logo for one 8×16 image block. The bits 1−52 are identical to bits 53−104. The remaining 24 bits are used for information about camera ID, original image resolution, etc.

6. CONCLUSION

In this paper, we analyze the security of fragile image authentication schemes based on watermarking that can accurately localize modifications. We propose a new, simple, and elegant scheme that resists all known attacks. We start with a comparison of authentication techniques that use cryptography and watermarking pointing out the main advantages and disadvantages of both. After a clear need for authentication based on watermarking is established, we proceed with classification of attacks on watermarking authentication techniques. We evaluate the security of current fragile authentication methods with respect to the attacks and outline a practical scenario for such schemes. It is argued that current schemes are vulnerable to attacks and do not provide a satisfactory and practical solution the problem with replacing blocks between images (the Holliman-Memon attack[19]). As a result, we formulate four design requirements that a secure and practical scheme should satisfy. The new scheme is based on the Wong's scheme[18] in which the logo contains information about the image index, block index, image dimensions, and other ancillary data. This idea enables us separate the authentication of the content and the content's origin. The logo structure is used to verify the block integrity while the logo content gives us the information about the block origin (block index, the image index, author ID, etc.). The new method is very fast, simple, and elegant, and it also appears to resist all five attacks introduced in Section 3, including the Holliman-Memon attack. In particular, it can be used in a practical scenario for authentication of images in digital cameras using either private or public key infrastructures.

ACKNOWLEDGMENTS

The work on this paper was supported by Air Force Research Laboratory, Air Force Material Command, USAF, under the grants F30602-99-2-0541 and F30602-98-C-0176. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation there on. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of Air Force Research Laboratory, or the U. S. Government. Special thanks belong to Miroslav Goljan for useful discussions in preparation of this paper.

REFERENCES

1. Celik, M., Sharma, G., and Saber, E.: A Hierarchical Image Authentication Watermark With Improved Localization And Security, In: Proc. ICIP 2001(CD ROM version), paper ID 3532, Thessaloniki, Greece, October (2001)
2. Coppersmith, D., Mintzer, F., Tresser, C., Wu, C. W., Yeung, M. M.: Fragile Imperceptible Digital Watermark with Privacy Control. In: Proc. SPIE, Security and Watermarking of Multimedia Contents, San Jose, California, January (1999) 79–84
3. Friedman, G. L.: The Trustworthy Digital Camera: Restoring Credibility to the Photographic Image. IEEE Transactions on Consumer Electronics **39** (4), November (1993) 404–408
4. Fridrich, J.: Visual Hash for Oblivious Watermarking. In: Proc. SPIE Security and Watermarking of Multimedia Contents, San Jose, California, January (2000) 286–294
5. Fridrich, J., Goljan, M., Memon, N.: Further Attacks on Yeung-Mintzer Fragile Watermarking Scheme. In: Proc. SPIE, Security and Watermarking of Multimedia Contents, San Jose, California, January (2000) 428–437
6. Fridrich, J., Goljan, M., Memon, N.: Cryptanalysis of the Yeung-Mintzer Fragile Watermarking Technique. Submitted to the *Journal of Electronic Imaging*, March (2000)
7. Fridrich, J., Goljan, M., Du, R.: Distortion-Free Data Embedding. The 4th Information Hiding Workshop, Pittsburgh, Pennsylvania, April (2001)
8. Fridrich, J., Goljan, M., Baldoza, A. C.: New Fragile Authentication Watermark for Images. In: Proc. ICIP, Vancouver, Canada, September (2000)
9. Holliman, M. and Memon, N.: Counterfeiting Attacks on Oblivious Block-Wise Independent Invisible Watermarking Schemes. In: Proc. IEEE Trans. on Image Processing **9** (3), March (2000) 432–441
10. Kundur, D. and D. Hatzinakos, D.: Towards a Telltale Watermarking Technique for Tamper Proofing. In: Proc. ICIP, Chicago, Illinois, October (1998)
11. Lin, C.-Y. and Chang, S.-F.: Semi-Fragile Watermarking for Authenticating JPEG Visual Content. In: Proc. SPIE, Security and Watermarking of Multimedia Contents, San Jose, California, January (2000) 140–151
12. Marvel, L. M., Hartwig, G. W., Boncelet, C. Jr.: Compression-Compatible Fragile and Semi-Fragile Tamper Detection. In: Proc. SPIE, Security and Watermarking of Multimedia Contents, San Jose, California, January (2000) 140–151
13. Memon, N., Shende, S., Wong, P.: On the Security of the Yeung-Mintzer Authentication Watermark. In: Proc. of the IS & T PICS Symposium, Savannah, Georgia, March (1999)
14. Schneider, M. and Chang, S.-F.: A Content-Based Approach to Image Signature Generation and Authentication. In: Proc. ICIP '96, Lausanne, Switzerland, October (1997) 227–230
15. Walton, S.: Information Authentication for a Slippery New Age. *Dr. Dobbs Journal* **20** (4), April (1995) 18–26
16. Wolfgang, R. B. and Delp, E. J.: A Watermark for Digital Images. In: Proc. ICIP vol. 3, Lausanne, Switzerland, September (1996), 219–222
17. Wolfgang, R. B. and Delp, E. J.: Fragile Watermarking Using the VW2D Watermark. In: Proc. SPIE, Security and Watermarking of Multimedia Contents, San Jose, California, January (1999) 204–213
18. Wong, P.: A Watermark for Image Integrity and Ownership Verification. In: Proc. IS&T PIC, Portland, Oregon (1998)
19. Wong, P. and Memon, N.: Secret and Public Key Authentication Watermarking Schemes that Resist Vector Quantization Attack. In: Proc. SPIE, Security and Watermarking of Multimedia Contents, San Jose, California, January (2000) 417–427
20. Xie, L. and Arce, G.: Joint Wavelet Compression and Authentication Watermarking. In: Proc. ICIP, Chicago, Illinois, October (1998) 427–431
21. Yeung, M. M. and Mintzer, F.: An Invisible Watermarking Technique for Image Verification. In: Proc. ICIP'97, Santa Barbara, California (1997)