NEW FRAGILE AUTHENTICATION WATERMARK FOR IMAGES

[a)]Jiri Fridrich, [b)]Miroslav Goljan, and [c)]Arnold C. Baldoza

[a)]Center for Intelligent Systems, SUNY Binghamton, Binghamton, NY 13902-6000
[b)]Department of Electrical Engineering, SUNY Binghamton, NY 13902-6000
[c)]First Lieutenant, United States Air Force
{fridrich, bg22976}@binghamton.edu, baldoza@concentric.net

ABSTRACT

In this paper, we propose a new fragile watermark for image authentication. Based on the Yeung-Mintzer scheme [1], the new watermark does not have certain security gaps common to many previously proposed fragile watermarks. A block cipher is used instead of binary look-up tables. Pixel values are perturbed by small quantities so that the cipher maps small pixel neighborhoods to a fixed binary logo. This process is further modified in order to embed image indices (time stamps) into disjoint blocks of every image. This is necessary for detection of collages from multiple authenticated images. We also formulate basic security requirements and investigate the security of the new scheme.

1. INTRODUCTION

The goal of authentication using cryptography is to protect the communication channel. The cryptographic protocol provides a guarantee that the image has not been tampered with and came from the right person. This is typically achieved using hash functions and a public-key cipher. The properties of hash functions prevent us from localizing or quantifying the changes that have been made to the image. For example, flipping one pixel will result in a tampered image even though the change would have virtually no impact on the visual quality of most images. The ability to localize where the changes have occurred is clearly a desirable feature. Digital watermarking provides a convenient tool for authentication of visual information, tamper detection, and verification of image integrity [1–4]. Watermarking as opposed to pure cryptographic tools enables us to localize tampered or damaged areas or even authenticate with a degree [5,6] without having to store any additional information about the image. With the exception of [7], this is typically achieved at the expense of introducing a slight amount of distortion into the image. Because the auxiliary authentication information is embedded in the image itself rather than appended, it gives us more flexible and convenient tool for investigating the image integrity.

One of the first watermarks proposed as a means for authentication were fragile watermarks. Walton [3] proposed to hide key-dependent check-sums of the seven most significant bits (MSBs) of grayscales along pseudo-random walks in the least significant bits (LSBs) of pixels forming the walk. Better and more secure schemes can be constructed using cryptographic hash functions. Wong [Won] described a scheme in which an image is divided into blocks and each block contains the hash calculated from the MSBs in the LSBs of the pixels forming that block. The size of the block is limited from below by the length of the hash. Thus, the localization properties of such a scheme are also limited. In addition to that, it is possible to swap blocks within one image and among different images without introducing detectable changes. Yeung and Mintzer [1] proposed an elegant scheme that authenticates individual pixels. This scheme has very good localization properties but, unfortunately, it is not easy to develop a feasible key management infrastructure without introducing security gaps [8–10]. Other fragile authentication watermarks and watermarks combined with hash functions were proposed in [4]. A highly original approach to image authentication has been recently proposed by Honsinger [7]. In his scheme, the process of authentication can be inverted if the image is indeed tested as authentic. This "invertible authentication" is possible only at the expense of not being able to authenticate all possible images. In practice, this is not a problem because all typical images can be authenticated.

In this paper, we propose a new fragile watermarking scheme that does not have certain weak points exhibited by a majority of previously proposed fragile authentication watermarks. In the development of the overall design of this scheme, we started with the well-known Yeung-Mintzer scheme and identified certain weak points when the same key and binary logo is reused for different images [10]. To eliminate the weak points, we replaced the binary look-up tables for a single pixel by an encryption map defined on a local neighborhood rather than on a single pixel. The image is authenticated by scanning the image by rows and modifying the pixel colors in the same manner as in the Yeung-Mintzer scheme. The algorithm can be combined with dithering and error diffusion for the best visual results. To prevent the collage attack described in [8, 10], in which portions of different images are pasted together, we embed an image index into each image repetitively into a set of non-overlapping 32×32 blocks. This is achieved by slightly modifying the first scanning step. In particular, we embed the image index in a binary form into a set of randomly pre-selected pixels by adding another condition their colors need to satisfy – the values of a different encryption-based look-up table should return the right index bit. The scheme is capable of detecting

collages from multiple images, and is not vulnerable to previously proposed attacks.

2. SECURITY REQUIREMENTS

We start this section with a set of observations: It is not enough to propose an authentication watermark that works reasonably well for one image. We have to explain in detail how a large number of images will be watermarked (databases, video-frames, or images taken with a digital camera). Do we use the same key for all images, or a different key for each image? If a different key is used for each image, how complex is the tool that checks the authenticity of images? Is there any leakage of information from the scheme if multiple pairs of original and authenticated images become available? What is the probability $P(n)$ of modifying n pixels without introducing detectable changes? Without investigating these questions, the authentication scheme is almost useless because it is not clear if it can be trusted. As explained in the introduction, authentication based on cryptography and hash functions is limited but it can provide guarantees and exact mathematical estimates of probabilities that a given image has been or has not been tampered. Authentication based on watermarking claims to be a better tool that can localize changes. However, this benefit should not jeopardize the security of the scheme. As pointed out in the papers by Memon and Fridrich [8−10], most watermark authentication schemes contain profound weaknesses that limit their usefulness. In this section, we start with specifying a set of requirements that a secure authentication scheme should posses. A similar list was for the first time put forward by Memon [9].

Given multiple pairs of original and authenticated images, the following issues should be addressed:

- Is there any information leakage from the scheme? In other words, is it possible to obtain some information about the secret key, locations of pixels carrying information, the logo, and other entities that enter the authentication scheme?
- Can we swap blocks/frequency bins/wavelet coefficients within the image and among different images without causing a detectable change? Surprisingly many schemes are vulnerable to this type of attack. In the original Yeung-Mintzer scheme, one could swap pixels with the same spatial location among different authenticated images without introducing any detectable changes. This attack has been described for the first time in [8] and further analyzed in [10]. It is not limited to block-based schemes, but can actually be applied to many other schemes that do not use blocks but other entities, such as frequency bins or wavelet coefficients for authentication.
- Is it possible to use denoising algorithms to extract an approximation to the watermark and use it to authenticate an arbitrary image? If the watermark is a strong, nontrivial function of the image, this attack will not be possible. Some watermarking schemes for copyright protection may be vulnerable to this type of attack [12].
- What is the probability of making a modification in the authenticated image that will go undetected? How does the probability increase with the spatial extent of the changes (number of tampered pixels)? This item is important if the scheme should ever be questioned in the court.
- Having the authentication detector available, can it be misused? Is it possible to create a forgery? Is it possible to modify an image locally so that the changes will not be detected?

3. NEW AUTHENTICATION SCHEME

Our scheme is based on the work of Yeung and Mintzer [1]. They proposed the following method for authentication of digital images. The process of image authentication starts with a secret key that is used to generate a key dependent binary valued function f, f: $\{0, 1, \ldots, 255\} \rightarrow \{0,1\}$, that maps integers 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 gray scales are perturbed to satisfy the following expression

$$L(i,j) = f_g(g(i,j)) \text{ for each pixel } (i,j).$$

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

$$L(i,j) = f_R(R(i,j)) \oplus f_G(G(i,j)) \oplus f_B(B(i,j)) \text{ for each pixel } (i,j).$$

Error diffusion is further employed to preserve the original colors. The image authenticity is easily verified by checking the relationship $L(i,j) = f_g(g(i,j))$ for each pixel (i,j).

Attack 1: It has been pointed out in [10] that if the same logo and binary function f are reused for multiple images, it becomes very easy to accurately estimate both the logo and the binary function f. The attack is a little more complicated for color images, but the security problem persists. Actually, color images provide even more constraints and fewer images are needed to recover the binary tables and the logo.

Attack 2: Another attack that can be mounted against any watermarking scheme in which the watermark is local and does not depend on the image index is based on combining portions from different images while keeping their relative positions within the image (creating a collage). This "collage" attack has been described for the first time in [8] and analyzed in detail in [10].

It appears that the simplest solution to prevent both attacks would be to make the fragile watermark depend on the image index (or the time stamp). This would solve the collage attack completely, and since the binary functions f and the logo L can also be made index-dependent, we could not apply the first attack either. However, now we need the index in order to verify the authenticity of an image. An exhaustive search may not be plausible because of a potentially large number of authenticated images. A better idea would be to embed the index in the image using a key uniquely associated with a particular digital camera or a movie. The index cannot be embedded in the whole image because tampering with a portion of the image might result in a conclusion that the whole image has been tampered with. Thus, the index needs to be embedded multiple times in small blocks. Also, it should be embedded in a robust manner so that the correct index is extracted even from slightly modified blocks. However, robust embedding would lead to larger distortion

contributing to the distortion due to the fragile watermark. Obviously, the robust watermark carrying the information about the index must be embedded as the first watermark. To achieve at least moderate robustness of the index watermark, we may have to increase the block size to at least 64×64 pixels. Assuming the index can be captured using 32 bits, we are taking a potentially large risk that tampering with a very spatially localized feature will cause an erroneous index extraction. Increasing the block size negatively influences the ability of the authentication scheme to localize changes.

In this paper, we propose the following approach. Instead of making the fragile watermark dependent on the image index, we keep it index-independent and propose to replace the binary look-up tables with a mapping derived from a block encryption function. We involve more than one pixel in the mapping to prevent the first attack. For example, assuming we are scanning the picture by rows, we can include neighbors from the portion of the image that has already been modified. This modification will prevent attacks based on calculating the lookup tables. To enable detection of collages, we propose to insert the image index in a fragile manner into randomly chosen pixels in small blocks by slightly modifying the authentication process. We propose the following scheme for grayscale images as it would be implemented for authentication of digital images taken with a digital camera:

1. The camera key K is used to generate a key for a secure block encryption algorithm E_K. We further choose an integer a (a=5) and a random walk through the $a \times a$ square (permutation of the set of integers $\{1, \ldots, a^2\}$).
2. The image is padded on the left edge and the upper edge by $a-1$ pixels with a key-dependent random signal. Then, the image is scanned by rows, starting in the upper left corner, progressing to the right and down.
3. For each pixel g_{ij}, we look at the $a \times a$ square with pixels $g_{i-u\ j-v}$, $0 \le u, v \le a-1$. It is the square that contains the pixel g_{ij} at its lower right corner. Let $g_1, \ldots, g_{a \times a}$ are the gray scales of the a^2 pixels in this square in the order of the random walk generated in Step 1.
4. The gray scale g_{ij} is modified by going through its neighboring shades till we obtain

$$L(i, j) = f_K(g_1, \ldots, g_{a \times a}) = Parity(E_K(g_1, \ldots, g_{a \times a})), \quad (1)$$

   where L is the binary logo, and $E_K(g_1, \ldots, g_{a \times a})$ is the encrypted bitstream. Note, that the grayscales need to be converted to a binary stream before the encryption function E_K can be applied. By *Parity* we understand the excluded OR of the encrypted bitstream.
5. Error diffusion can be further applied to make sure that authentication will not introduce consistent error.

The first attack will not be successful for this scheme because of a large number of possible combinations of values of the binary function f_K. It is also not known in which order the pixels enter the encryption function E_K (there are $a^2!$ possible orderings of the pixels). On top of this, the secure block encryption algorithm provides additional feature of security making the first attack impossible to mount.

To enable detection of collages, we need to embed the image index into the image. We choose to embed this index into every 32×32 block into randomly chosen pixels (the selection of pixels depends on the secret key and stays the same for every block and every image). We can modify Step 4 by including another constraint whenever we are at a pixel that carries information about the index. For example, using a different key K', we can add the following constraint (at index-carrying pixels (i, j) only):

$$IB(i, j) = Parity(E_{K'}(g_1, \ldots, g_{a \times a})), \quad (2)$$

where $IB(i, j)$ denotes the required image index at the index-carrying pixel (i, j).

During the detection, we first apply the detection function for the fragile watermark and by correlating with the logo, we can identify portions of the image that have been tampered with. We can also determine if the image has been cropped. If an attacker combines segments from different images but keeps their relative positions, we only detect tampering at the boundaries of those segments. To establish whether or not the segments came from different images, we recover the image indices from each 32×32 block.

4. SECURITY ANALYSIS

Looking back at the requirements set forth in Section 2, we analyze the security of the new scheme. If one pixel in an image is randomly perturbed, the probability that this modification will go undetected is $1:2^{a \times a}$ because the pixel influences the relationship (1) for $a \times a$ pixels to the right and down. If the pixel is near the right or the bottom boundary of the image, the probability increases up to 1:2 for the pixel in the bottom right corner.

Given a set of original and authenticated images, one can build a database of $a \times a$ squares that are mapped to the same bit by the function f_K. If we collected all possible $a \times a$ squares, we could easily authenticate any image even without knowing the key. However, the number of different $a \times a$ squares is too high ($256^{a \times a}$) for this attack to be practical. An incomplete database will not enable us to create any large images (i.e., images larger than $a \times a$).

We acknowledge that unlimited access to the detector and its output bitmap (pixels matching the logo bits) is equivalent to having access to the authenticating mechanism. One could simply scan the image by rows and modify the pixels till they become authenticated. This would require on average $M \times N$ accesses to the detector. This limits the usability of the scheme to situations where the access to the detector is limited to trusted subjects, such as courts, or law enforcement.

5. ACKNOWLEDGEMENTS

The work on this paper was supported by Air Force Research Laboratory, Air Force Material Command, USAF, under a grant number F30602-98-C-0009. 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.

6. REFERENCES

[1] M. Yeung, and F. Mintzer, "An Invisible Watermarking Technique for Image Verification", *Proc. ICIP'97*, Santa Barbara, California, 1997.

[2] P. Wong, "A Watermark for Image Integrity and Ownership Verification", *Proc. IS&T PIC Conference*, Portland, Oregon, 1998.

[3] S. Walton, “Information Authentication for a Slippery New Age”, *Dr. Dobbs Journal*, vol. 20, no. 4, pp. 18–26, Apr 1995.

[4] R. B. Wolfgang and E. J. Delp, "Fragile Watermarking Using the VW2D Watermark", *Proc. SPIE, Security and Watermarking of Multimedia Contents*, San Jose, California, Jan 25–27, 1999, pp. 204–213.

[5] J. Fridrich, "A Hybrid Watermark for Tamper Detection in Digital Images", *ISSPA'99 Conf.*, Brisbane, Australia, August 22–25, 1999.

[6] D. Kundur and D. Hatzinakos, "Towards a Telltale Watermarking Technique for Tamper Proofing", *Proc. ICIP*, Chicago, Illinois, Oct 4–7, 1998, vol 2.

[7] C. W. Honsinger, P. Jones, M. Rabbani, J. C. Stoffel, "Lossless Recovery of an Original Image Containing Embedded Data", US Patent application, Docket No: 77102/E–D, 1999.

[8] N. Memon, S. Shende and P. Wong. “On the security of the Yeung-Mintzer Authentication Watermark. *Proc. of the IS & T PICS Symposium*, Savannah, Georgia, March 1999.

[9] M. Holliman and N. Memon. “Counterfeiting attacks for block-wise independent watermarking techniques.” *To.appear in IEEE Transactions on Image Processing*, March 2000.

[10] J. Fridrich, M. Goljan, N. Memon, "Further Attacks on Yeung-Mintzer Watermarking Scheme", *Proc. SPIE Electronic Imaging 2000*, San Jose, January 24–26, 2000.

[11] M. Holliman, N. Memon, and M. M. Yeung, "On the Need for Image Dependent Keys for Watermarking*", Proc. Content Security and Data Hiding in Digital Media*, Newark, NJ, May 14, 1999.

[12] M. Kutter*,* "Watermark Copy Attack", *Proc. SPIE Photonic West Electronic Imaging 2000, Security and Watermarking of Multimedia Contents,* San Jose, California, January 24–26, 2000.