# A SECURE IMAGE AUTHENTICATION ALGORITHM WITH PIXEL-LEVEL TAMPER LOCALIZATION*

*Jinhai Wu[1], Bin B. Zhu[2], Shipeng Li[2], Fuzong Lin[1]*
[1]State Key Lab of Intelligent Tech. & Systems, Tsinghua Univ., Beijing, 100084, P. R. China
[2]Microsoft Research Asia, Beijing, 100080, P. R. China

**ABSTRACT**

*In this paper, we use a generalized model for all the previously proposed pixel-wise authentication schemes. Then we show how this model can be compromised with an oracle attack. This motivated us to develop a novel authentication scheme to be described here. The new scheme is designed to detect any changes to a signal. It consists of two mechanisms dedicated to authenticity verification for a signal and localization of tampered pixels, respectively. The scheme is secure yet maintains the fine tamper localization capability of a pixel-wise scheme. Experimental results show that the proposed scheme has a good accuracy in locating tampered pixels.*

## 1. INTRODUCTION

Multimedia authentication is a technology to check authenticity and integrity of multimedia signals. It is often desirable to localize tampered pixels or samples for a tampered signal so unmodified parts can still be used. Technologies to fulfill this goal have been actively studied in recent years. A class of proposed technologies, called complete or hard authentication, is to detect any modifications to a multimedia signal. Hard authentication technologies can be classified into pixel-wise schemes and block-wise schemes. A pixel-wise scheme is designed to localize tampered pixels (or samples for audio signals) in additional to verify authenticity for the whole signal. A block-wise scheme, on the other hand, is designed to localize tampered blocks. A block-wise scheme is securer in general than a pixel-wise one, but has much coarser tamper localization capability. Details on proposed authentication technologies can be found in [1][2].

One of the first pixel-wise authentication schemes was the fragile watermarking scheme proposed by Yeung and Mintzer (Y-M scheme) [3][4]. For grayscale images, the Y-M scheme applies a secret binary function to map the value of each pixel, perturbed if necessary, to a preset logo bit. The scheme is able to localize a single tampered pixel. Its vulnerabilities under various circumstances were reported in [5]-[9], and fixes in [10]-[13]. A typical fix is to introduce neighborhood dependency in mapping a pixel to a logo bit, such as the scheme proposed in [10]. These fixes can thwart the attacks reported in [5]-[9], but, as pointed out by Fridrich in [14], are vulnerable to oracle attacks if the pixel scan order, i.e., the order that pixels are watermarked in the embedding process, is public, and if the oracle returns locations of the detected tampered pixels. Fridrich attributed this new vulnerability to the inherent sequential nature in modifying pixels during the watermarking process in a pixel-wise scheme, and believed that no pixel-wise schemes could fix this vulnerability. She turned attention to develop a block-wise scheme in [14] which does not suffer from any of the aforementioned vulnerabilities for pixel-wise schemes. Unfortunately, a block-wise scheme greatly reduces the tampering localization capability. A tampered pixel can no longer to be identified.

In this paper, we re-examine oracle attacks with a generalized model for all reported pixel-wise schemes. We find out that all these schemes, no matter the pixel scan order is public or secret, are vulnerable under oracle attacks. This agrees with Fridrich's conclusion about pixel-wise authentication schemes in [14]. A close study turns out that the real cause of the problem is that all the proposed pixel-wise schemes use a single mechanism for two very different purposes: authenticity verification and tamper localization. We then propose a secure multimedia authentication scheme with dual mechanisms, one is dedicated to authenticity verification, and the other is dedicated to tamper localization. This novel scheme is secure to all the known attacks yet maintains pixel-level tamper localization.

The rest of this paper is organized as follows: Oracle attacks for a generalized model of proposed pixel-wise schemes are described in Section 2. Our new authentication scheme is then described in Section 3. Experimental results are reported in Section 4 before we conclude the paper in Section 5.

## 2. ORACLE ATTACKS

### 2.1. Generalized Model of Pixel-Wise Schemes

All the pixel-wise schemes proposed in [3][4][10]-[13] watermark pixels sequentially, one pixel at a time. They can be described by the following generalized model: an image is scanned in a certain order to map a 2-D image to a 1-D vector. The scan order may be public such as in most proposed pixel-wise schemes or secret. For example,

* This work was done when Jinhai Wu was a visiting student at Microsoft Research Asia.

0-7803-8554-3/04/$20.00 ©2004 IEEE.

the scheme proposed in [10] embeds pixels in a row-by-row scan order, and the scheme in [11] in a zig-zag scan order. Pixels are modified in the watermarking process according to the order in this 1-D vector. A neighborhood-dependent mapping function is used to map each pixel value, perturbed if necessary, to a desired logo bit (for simplicity, grayscale images are used here but the results are also valid for color images). The neighborhood of a pixel does not necessarily mean the actual neighborhood in the original image or the 1-D vector. It is a generalized term to mean some set of previously processed pixels. Preset values may be added to the 1-D vector to be used as neighborhoods for some image pixels, typically the pixels near the beginning of the 1-D vector. Note that the neighborhood for the Y-M scheme is empty, i.e., no neighborhood is used.

### 2.2. Oracle Attack to Generalized Model

An authenticated image and an oracle are needed in performing an oracle attack. The image is modified and then sent to the oracle which returns the verification result, i.e., if the testing signal is authentic or not. We assume that the oracle also returns the locations of detected tampered pixels. An efficient oracle attack to compromise the Y-M scheme was proposed in [9]. For all the pixel-wise schemes, another oracle attack when the scan order is public was briefly mentioned by Fridrich in [14]. In this section, we describe an oracle attack to the aforesaid generalized model of pixel-wise schemes. The scan order can be secret.

Our oracle attack consists of two stages. The first stage deduces the secret scan order. This stage can be skipped if the scan order is known. The second stage modifies an authentic image with any desired content yet passes the authenticity verifier. In the first stage, each pixel is randomly modified and the locations of tampered pixels returned by the oracle are recorded. All these tampered pixels are scanned, i.e., placed in the 1-D vector after the current modified pixel. One pixel is modified at a time. For example, if pixel $A$ is used as a neighboring pixel in watermarking pixel $B$, then a change to $A$ has 50% probability on average to make $B$ reported as tampered. Hence, to modify A twice on average can find all the pixels, no matter how many, which use A as one of the neighboring pixels. Therefore $2MN$ oracle tests on average are needed for a grayscale image of $M \times N$ pixels to deduce the secret scan order. We note that it is possible the deduced 1-D vector might be different from yet equivalent to the 1-D vector used in watermarking, if a scheme has several equivalent scan orders. For example, if a scheme partitions an image into two disjoint subimages which are watermarked independently, then it does not matter which subimage is scanned (i.e. watermarked) first.

The second stage generates an approximation of an arbitrary desired image $I^d$ of the same size as the authenticated image $I$ yet tested authentic. Suppose the deduced scan order in the first stage is $S$. The index $P$ of the current scan position is initialized to 0, the beginning of the scanned 1-D vector. The resulting image $I^r$ is initialized to $I^d$. The following procedure is applied:

1. Starting from the current position $P$, the pixels in $I^r$ and $I$ are compared according to the scan order $S$ to find the first unmatched pixel $C \in I^r$. This stage is terminated if no such a pixel can be found.
2. Repeat testing the image with the pixel $C$ perturbed until $C$ is returned as untampered by the oracle. Set $I^r$ to be the perturbed image. If there is any tampered pixel returned by the oracle, set $P$ to be the first tampered pixel according to the scan order $S$ and go to Step 1. Otherwise terminate this stage.

The resulting image $I^r$ is an approximation of $I^d$ that we are seeking for. It is apparent that $I^r \neq I$ yet is verified authentic. The number of oracle tests in the second stage is about $2MN$ or less on average. The perturbation in Step 2 above can be done in the same way as in the watermarking process so $I^r$ is expected to be perceptually close to $I^d$. In fact, the second stage is very similar to the watermarking process.

The total number of oracle tests in the described oracle attack is about $4MN$ when the scan order is secret and about $2MN$ when the scan order is public. This oracle attack works for color images in a similar way.

## 3. OUR WATERMARKING SCHEME

In realizing oracle attacks to pixel-wise schemes, Fridrich believed that block-wise schemes were the only viable solution and proposed a block-wise authentication scheme which can localize tampered blocks rather than a single pixel [14]. For security reason, the size of a block should be 128 pixels or larger. It is desirable in many applications to have a finer tamper localization capability. A natural question arises: Is it possible to develop a secure authentication scheme with a tamper localization capability as fine as pixel-wise schemes?

For the generalized model described in Section 2.1, an image is claimed authentic if no pixel is found tampered. The authenticity of a pixel is checked by applying a many-to-one mapping function to map the value of each pixel to a bit which is compared against a logo bit. Pixels are watermarked sequentially, one pixel at a time. These features enable pixel-wise schemes with good perceptual quality, but are also exploited by oracle attacks. If we

separate signal authentication verification from tamper localization, and make pixel-wise schemes dedicated to tamper localization, then we should be able to develop a secure authentication scheme with pixel-level tamper localization capability. This is the core idea for the novel scheme to be described in the rest of this section.

### 3.1. Our Authentication Scheme

Our fragile watermarking scheme merges the bests of a pixel-wise scheme and of a block-wise scheme: A pixel-wise mechanism is dedicated to localizing tampered pixels and a block-wise mechanism is dedicated to authenticity verification for a signal. The latter mechanism treats the whole image as a block. To improve detection probability of a single tampered pixel, the former mechanism uses a previously watermarked pixel in the secret pixel scan order as the neighborhood of the current pixel. Each mechanism can be optimized for its assigned task. Due to the length limitation of this paper, 8-bit grayscale images are assumed in describing our scheme.

**Watermark embedding:** The watermarking procedure for a grayscale image $I$ is as follows:

1.1. A secret key $K$ is used to generate a random mapping function $f$ which maps an integer in [0, 255] to 0 or 1.

1.2. $K$ is used to shuffle $I$ to a randomized image $X = Shuffle_K(I)$. Both $X$ and the binary logo $L$ are then ordered into 1-D vectors of length $N$ by either zigzag or row-by-row scan, where $N$ is the number of pixels in the image $I$.

1.3. All the pixels in $X$ are partitioned into two disjoint subspaces $A$ and $B$, where $B$ contains the last $r$ pixels and $A$ contains the rest pixels. The least significant bit (LSB) of each pixel in $B$ is set to zero. The value of $r$ is discussed in Step 1.5.

1.4. For $i$-th pixel $X(i)$, where $i$ runs from 1 to $N$, the following relationship is enforced with $X(i)$ possibly perturbed if necessary:

$$L(i) = f(X(i-1) \oplus X(i)), \qquad (1)$$

where $X(0) \equiv 0$ and $\oplus$ stands for XOR operation. If $X(i)$ is in $B$, the perturbed value should be even, i.e., the LSB after perturbation is still 0.

1.5. The result is hashed with MD5 and encrypted by the private key of an asymmetric encryption to generate a digital signature $D$. Alternatively, a keyed hash or MAC can be used to generate $D$. The value of $r$ in Step 1.3 is the size of $D$ in number of bits. $D$ is embedded into the LSBs of the pixels in $B$.

1.6. Reverse the scan and the shuffling operation in Step 1.2 to obtain the watermarked image.

**Authenticity verification:** Authenticity of a challenged image $I'$ is verified as follows:

2.1. Use the secret key $K$ to generate the same secret mapping function $f$ and get $X' = Shuffle_K(I')$. $X'$ is then ordered and the subspaces $A'$ and $B'$ of $X'$ are found in the same way as in Step 1.1.

2.2. The embedded $D'$ is extracted from the LSBs of the pixels in $B'$, decrypted if necessary, to get the original hash value $H$. The LSB of each pixel in $B'$ is set to 0.

2.3. Apply the same method as in Step 1.5 to the resulting $X'$ to obtain a hash value $H'$. If $H' = H$, then the challenged image is claimed authentic; otherwise tampered.

**Tamper localization:** For an inauthentic image, tampered pixels can be localized as follows:

3.1. Eq. (1) is applied to the result of Step 2.2 to extract the embedded logo $L'$. Scan $L'$ and $L$ into 1-D vectors, and find the set $S_D = \{i \mid L'(i) \neq L(i)\}$, which is then expanded as $S = S_D \cup \{i-1 \mid i \in S_D\}$. If $S$ is empty, the tampered pixels cannot be localized by the scheme and the procedure terminates.

3.2. Reverse the scan and the shuffling operation to find out the set of pixels $S^*$ corresponding to the set $S$. Those pixels in $S^*$ are potentially tampered.

Tamper localization can be further refined by exploiting the fact that typical manipulations in real applications result in connected modified pixels. The pixels in $S^*$ that are isolated or whose connected paths are smaller than a preset threshold are removed from $S^*$. The remaining pixels in $S^*$ are claimed to be tampered pixels. It appears that the tamper localization of the proposed scheme is very close to the Y-M scheme for typical manipulations in real applications, yet with higher detection probability of a tampered pixel, thanks to the neighborhood dependency used in our scheme.

### 3.2. Security Analysis

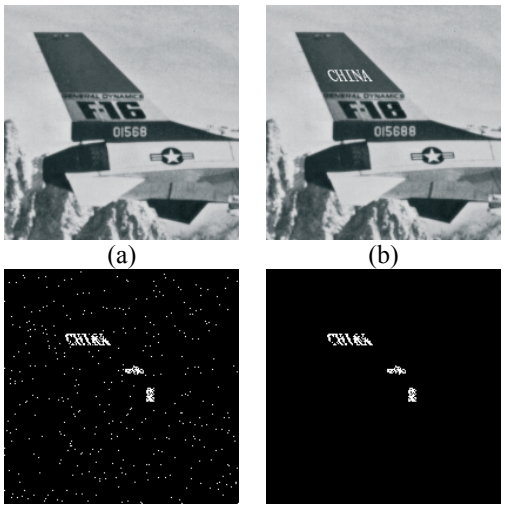
In our scheme, a cryptographic hash or MAC function is used to generate a digest for an image except the LSBs of the $r$ pixels used to embed the digest or its encrypted version. Any change to an authenticated image will be detected by the authenticity verification procedure, except when unlikely collision of the cryptographic hash or MAC function occurs. It is therefore impossible for an attacker to successfully launch the oracle attack described in Section 2.2 or any other known attacks. On the other hand, it is possible that some tampered pixels can not be detected by the tamper localization procedure. In fact, the proposed scheme has about $1-0.5^2=75\%$ probability to localize a tampered pixel, as compared to 50% detection probability in the Y-M scheme. Higher detection

probability can be achieved by increasing the number of neighborhood pixels used in Eq. (1) at the cost of reduced tamper localization capability.

If keyed hash or MAC is used in Step 1.5, and if the number of mismatched bits when $H$ and $H'$ in Step 2.3 are compared is much smaller than half of the hash bits, we can conclude that LSBs of the pixels in the subspace $B'$ corresponding to the mismatched bits are manipulated. When asymmetric encryption is used in Step 1.5, we can no longer make such a conclusion. The number of bits to be embedded is also increased. The advantage is that authorized detectors cannot forge a digital signature.

## 4. EXPERIMENTS

The proposed scheme was implemented and tested with images. The watermarked image of a plane is shown in Fig. 1.(a). This image was then manipulated by adding "CHINA" to the tail. The string of digits was changed from "01568" to "015688". The mark "F16" was also modified to "F18". The result of the manipulations is shown in Fig. 1.(b). Our eyes could not find any trace of manipulations. The proposed scheme was then applied to this manipulated image. The pixels in $S^*$, i.e., the detected tampered pixels without the post-processing refinement are shown in Fig. 1.(c). We can see a lot of isolated pixels spread over the whole image. The reported tampered pixels by our scheme, after the post-processing refinement, are shown in Fig. 1.(d)., which indicates the actual tampered pixels rather accurately.




(a) (b)

(c) (d)

Fig. 1 Experimental Results

## 5. CONCLUSION

We have presented a generalized model of previously proposed pixel-wise authentication schemes, and described in detail an oracle attack to compromise all these schemes. A secure authentication scheme with tamper localization as fine as a single pixel was then proposed. Experimental results showed that the tamper localization of the proposed scheme was rather accurate.

## REFERENCE


[1] B. B. Zhu, M. D. Swanson, and A. H. Tewfik, "When Seeing Isn't Believing," *IEEE Signal Processing*, vol. 21, no. 2, pp. 40-49, March 2004.

[2] B. B. Zhu and M. D. Swanson, "Multimedia Authentication and Watermarking," *Multimedia Information Retrieval and Management*, D. Feng, W. C. Siu, and H. Zhang, Eds., Springer-Verlag, Berlin, Heidelberg, New York, 2003, chap. 7, pp. 148-177.

[3] M. M. Yeung and F. C. Mintzer, "An Invisible Watermarking Technique for Image Verification," *IEEE Int. Conf. Image Processing*, 1997, vol. 2, pp. 680-683.

[4] M. M. Yeung and F. C. Mintzer. "Invisible Watermarking for Image Verification," *J. Electronic Imaging*, vol. 7, no. 3, pp. 578-591, July 1998.

[5] N. Memon, S. Shende, and P. Wong, "On the Security of the Yeung-Mintzer Authentication Watermark," *Proc. IS&T PICS Symp.*, Savannah, Georgia, March 1999, pp. 301-306.

[6] J. Fridrich, M. Goljan, and N. Memon, "Further Attacks on Yeung-Mintzer Fragile Watermarking Scheme," *Proc. SPIE vol. 3971 Security and Watermarking of Multimedia Contents II*, San Jose, CA, Jan. 2000, pp.428–437.

[7] M. Holliman and N. Memon, "Counterfeiting Attacks on Oblivious Block-wise Independent Invisible Watermarking Schemes," *IEEE Trans. Image Processing*, vol.9, no.3, March 2000, pp. 432-441.

[8] J. Fridrich, M. Goljan, and N. Memon, "Cryptanalysis of the Yeung-Mintzer Fragile Watermarking Technique," *J. Electronic Imaging*, vol. 11, pp.262-274, 2002.

[9] J. Wu, B. Zhu, S. Li, and F. Lin, "Efficient Oracle Attacks on Yeung-Mintzer and Variant Authentication Schemes," *IEEE Int. Conf. Multimedia & Expo*, Taiwan, Jun 2004.

[10] J. Fridrich, M. Goljan, and A. C. Baldoza, "New Fragile Authentication Watermark for Images," *IEEE Int. Conf. Image Processing*, Vancouver, Canada, Sept., 2000, vol. 1, pp. 446-449.

[11] C. T. Li, F. M. Yang, and C. S. Lee, "Oblivious Fragile Watermarking Scheme for Image Authentication," *IEEE Int. Conf. Acoustics, Speech, & Signal Processing*, Orlando, FL, USA, May 2002, vol. VI, pp 3445-3448.

[12] H. Zhong, F. Liu, and L. C. Jiao, "A New Fragile Watermarking Technique for Image Authentication," *Int. Conf. Signal Processing*, Aug. 2002, Beijing, vol. 1, pp. 792-795.

[13] H. Lu, R. Shen, and F. Chung, "Fragile Watermarking Scheme for Image Authentication," *Electronics Letters*, vol. 39, no.12, June 2003, pp. 898-900.

[14] J. Fridrich, "Security of fragile authentication watermarks with localization," *Proc. SPIE vol. 4675, Security and Watermarking of Multimedia Contents IV*, Jan. 2002, pp. 691-700.