Tampering with a watermarking-based image authentication scheme

This item was submitted to Loughborough University’s Institutional Repository by the/an author.

Citation: PHAN, R.C.-W., 2008. Tampering with a watermarking-based image authentication scheme. Pattern Recognition, 41 (11), pp. 3493-3496

Metadata Record: https://dspace.lboro.ac.uk/2134/8160

Version: Accepted for publication

Publisher: © Elsevier

Please cite the published version.
This item was submitted to Loughborough’s Institutional Repository (https://dspace.lboro.ac.uk/) by the author and is made available under the following Creative Commons Licence conditions.

For the full text of this licence, please go to:
http://creativecommons.org/licenses/by-nc-nd/2.5/
Tampering with a Watermarking-based Image Authentication Scheme

Raphael C.-W. Phan

Electronic & Electrical Engineering,
Loughborough University,
LE11 3TU, United Kingdom
Tel: +44 1509 227 080
Fax: +44 1509 227 014
r.phan@lboro.ac.uk

Abstract. We analyse a recent image authentication scheme designed by Chang et al. [Chang, C.-C., Hu, Y.-S., Lu, T.-C., 2006, A watermarking-based image ownership and tampering authentication scheme. Pattern Recognition Letters. 27, 439-446] whose first step is based on a watermarking scheme of Maniccam and Bourbakis [Maniccam, S.S., Bourbakis, N., 2004, Lossless compression and information hiding in images. Pattern Recognition. 37, 475-486]. We show how the Chang et al. scheme still allows pixels to be tampered, and furthermore discuss why its ownership cannot be uniquely binding. Our results indicate that the scheme does not achieve its designed objectives of tamper detection and image ownership.

Keywords: digital rights management, image authentication; watermarking; tampering; ownership

1 Introduction

An image authentication scheme allows to detect if any tampering has been performed on an image. Examples of watermarking-based [1–3, 5–7, 9–12, 15–20, 22–25] image authentication schemes are in [1–3, 5, 6, 10, 18, 22, 23, 25]. These types of schemes typically embed a watermark into the image that is a function of the image itself. For authentication, the watermark is recomputed and checked against the embedded one, thus any changes of the image will not pass the authentication check.

Chang et al. [5] recently proposed a watermarking-based image authentication scheme that is aimed to be secure against tampering. Its first step is derived from a SCAN-based [13] watermarking scheme by Maniccam and Bourbakis [14]. The difference is that the Chang et al. scheme does not employ SCAN patterns and makes use of a cryptographic hash function for feature extraction of the image blocks.

In this paper, we show how this scheme can be tampered and also discuss why its ownership is not uniquely binding. Our results disprove the security claims of
the scheme and conclude that it is not suitable for its designed purpose of image authentication and rightful ownership.

2 The Image Authentication Scheme

The first step of the image authentication scheme by Chang et al. is derived from that of the information hiding scheme of Maniccam and Bourbakis [14]: this step is used to decide on the number of embedded bits \( r \) for each pixel block.

The watermark embedding process extracts features of the image blocks and then for each block it embeds a function of its feature as a watermark into the middle pixel that represents that block. It is defined as follows:

A1. A grayscale image \( I \) of \( M \times N \) pixels is divided into \( M/2 \times N/2 \) overlapping blocks of size \( 3 \times 3 \) pixels. The center of each block is the watermarked pixel \( p^x \) (where \( x \) is the block index) used for watermarking in each block.

A2. For each watermarked pixel \( p^x \), the extracted block feature is actually the values of its eight-neighbours (as in standard computer graphics terminology) denoted \( \langle p^x_1, \ldots, p^x_8 \rangle \); plus private information specific to the image namely the block index \( x \), image identification \( ID \) and the image owner’s secret key \( SK \). See Fig. 1 for an illustration of a \( 3 \times 3 \) block, formed by the dotted boxes and \( p^x \) in the middle.

To be precise, the block feature is computed as

\[
\langle b^x_1, \ldots, b^x_{128} \rangle = H(p^x_1 || \ldots || p^x_8 || x || ID || SK),
\]

where \( H \) is a cryptographic hash function outputting 128 bits, and || denotes concatenation.

A3. The block variation \( \sigma \) is next computed as

\[
\sigma = \sum_{i=1}^{8} (p^x_i - p^x_{i+1 \mod 8})^2.
\]

A4. \( r \) is then determined as follows:

\[
r = \begin{cases} 
2; & 0 \leq \sigma < 8 \\
3; & 8 \leq \sigma < 16 \\
4; & 16 \leq \sigma < 255 
\end{cases}
\]

A5. The 128-bit block feature \( \langle b^x_1, \ldots, b^x_{128} \rangle \) is then padded with zeroes to a length of 132 bits, and folded (compressed) into \( r \) bits as follows:

\[
f^x = \bigoplus_{i=0}^{(132/r) - 1} (b^x_{ir+1} || \ldots || b^x_{ir+r}),
\]

where \( \oplus \) denotes exclusive-or sum.
A6. $f^x$ is then inserted into the $r$ least significant bits of the watermarked pixel $p^x$ to obtain the watermarked pixel $\hat{p}^x$. The image $I$ with its watermarked pixels $p^x$ replaced with the watermarked pixels $\hat{p}^x$ forms the final watermarked image $\hat{I}$.

The image authentication process checks if the watermarked image $\hat{I}$ has been tampered by recomputing the folded block features $f^x$ as per equation (3) and comparing them against the features $f^x$ that had been embedded within the watermarkable pixels $p^x$. It proceeds as follows:

B1. Steps (A1) through (A5) as defined above are performed, resulting in the recomputed block features $f^x$.

B2. For each block, $f^x$ is compared with the $f^x$ that had been embedded in $p^x$ to form the watermarked pixels $\hat{p}^x$ of the watermarked image $\hat{I}$. Equality means the block has not been tampered, otherwise it will be marked as tampered.

2.1 Security Claims

To be precise, we list here the claims made by Chang et al. for their image authentication scheme. In the next section we will demonstrate how these claims can be disproved.

**Claim 1 (Tamper Resistance).** The probability that each tampered block fails to be detected is

$$\left(\frac{1}{2}\right)^r ; \quad 2 \leq r \leq 4.$$

**Claim 2 (Rightful Ownership).** Only the person who owns the secret key $SK$ can prove the rightful ownership of the watermarked image; i.e. rightful ownership fails if the adversary correctly guesses $SK$, which on average should occur only with probability $2^{-k}$ where $k$ is the bit length of the secret key $SK$.

3 Tampering with the Scheme

3.1 Tampering the Watermarked Pixels

Recall that the Chang et al. scheme divides the image $I$ of $M \times N$ pixels into $M/2 \times N/2$ overlapping blocks of size $3 \times 3$ pixels.

For instance, a $7 \times 7$ pixelated image will be divided into $3 \times 3$ (i.e. 9) blocks of $3 \times 3$ pixels. See Fig. 1 where each pixel is represented by a box. For the first block ($x = 1$), its watermarkable pixel $p^1$ is right in the middle, and the 8-neighbours of $p^1$ are indicated by boxes with dots in them. Block 1 is a block of $3 \times 3$ pixels formed by the dotted boxes and the $p^1$ box.

**Proposition 1 (Breaking the Tamper Resistance).** There exists an attack that breaks the tamper resistance of the Chang et al. scheme with probability 1, therefore disproving Claim 1.
Proof. The attack is as follows:

C1. Compute \( r \) as per step (A4) of the watermark embedding process.

C2. For each watermarked pixel \( \tilde{p}^x \), tamper with the \( 8 - r \) most significant bits of this pixel to obtain the tampered watermarked pixel \( \tilde{p}^x \).

To see why the attack breaks the tamper resistance claim, we walk through the steps of the image authentication process. Let the untampered watermarked pixel be denoted as \( \hat{p}^x \).

B1. Recall that this step performs steps (A1) through (A5). Details are as follows:

A1. This step is performed on the tampered image to determine the blocks, including the tampered watermarked pixels \( \tilde{p}^x \) within each block.

A2. This step extracts the block features \( \langle b_1^x, \ldots, b_{128}^x \rangle \) as per equation (1). The key observation here is that this equation does not depend on the watermarked pixel (whether tampered or untampered, thus neither \( \hat{p}^x \) nor \( \tilde{p}^x \)). Hence, the block features \( \langle b_1^x, \ldots, b_{128}^x \rangle \) computed by this step for the tampered watermarked pixel \( \tilde{p}^x \) will equal the block features computed for the untampered watermarked pixel \( \hat{p}^x \).

A3. Similarly, the computation of the block variation \( \sigma \) does not depend on the watermarked pixel, hence it is the same for both the tampered watermarked pixel \( \tilde{p}^x \) and the untampered watermarked pixel \( \hat{p}^x \).

A4. Since \( r \) is a function of \( \sigma \), again the same value is computed for both the tampered and untampered watermarked pixels.

A5. Denote by \( f^{xz} \) the recomputed value as per equation (3). Since this equation is a function of the block features \( \langle b_1^x, \ldots, b_{128}^x \rangle \), hence the \( f^{xz} \) that is recomputed at this step for the tampered watermarked pixel \( \tilde{p}^x \) is
the same as the \( f_x \) computed for the untampered watermarkable pixel \( \tilde{p}_x \) and embedded into the watermarkable pixel during the **watermark embedding** process.

B2. Since the recomputed \( f_x' \) equals the \( f_x \) embedded in the watermarked pixel \( \tilde{p}_x \), the authentication check passes, and the block is marked as untampered. This occurs with probability 1; significantly higher than the probability of \( \left( \frac{1}{2} \right)^r \) in Claim 1.

\[ \Box \]

In summary, for any image of \( M \times N \) pixels, we can modify the \( 8 - r \) most significant bits of \( \frac{M}{2} \times \frac{N}{2} \) pixels without being detected, giving a total of \( \frac{(8 - r)MN}{4} \) bits.

### 3.2 Tampering the Unwatermarked Pixels

We now show how the unwatermarked pixels (represented as dotted boxes in Fig. 1) of a watermarked image can be tampered with while still passing the authentication check, and how it leads to breaking the rightful ownership claim of the Chang et al. scheme.

**Proposition 2 (Breaking the Rightful Ownership.)** There exists an attack that breaks the rightful ownership of the Chang et al. scheme with probability \( 2^{-r} (< 2^{-k}) \), therefore disproving Claim 2.

**Proof.** The attack is as follows:

D1. Compute \( r \) as per step (A4) of the **watermark embedding** process.

D2. For each block \( x \) of the watermarked image \( \hat{I} \), tamper with the 8 unwatermarked pixels \( \langle \hat{p}_{x1}; \ldots; \hat{p}_{x8} \rangle \) to obtain the tampered pixels \( \langle \tilde{p}_{x1}; \ldots; \tilde{p}_{x8} \rangle \). Denote the resultant tampered watermarked image as \( \tilde{I} \).

D3. Make \( 2^{-r} \) guesses \( \tilde{SK} \) of the secret key value, compute the corresponding block feature \( \tilde{f}_x \), and check for a match with the block feature \( f_x \) embedded in the watermarked pixel \( \tilde{p}_x \). On average, one match is expected, and this is the \( \tilde{SK} \) that is used by the adversary to lay ownership claims.

To see why this breaks the rightful ownership claim, we walk through the steps of the **image authentication** process. Denote the untampered unwatermarked pixels as \( \langle \hat{p}_{x1}; \ldots; \hat{p}_{x8} \rangle \).

B1. Recall that this step performs steps (A1) through (A5). More details are as follows:

A1. This step is performed on the tampered image \( \tilde{I} \) to determine the \( 3 \times 3 \) blocks, where each block has the watermarked pixel \( \hat{p}_x \) and its eight-neighbours \( \langle \hat{p}_{x1}; \ldots; \hat{p}_{x8} \rangle \) which have been tampered with.

A2. With the secret key \( \tilde{SK} \) supplied by the adversary computed from step (D3) above, this step extracts the block features \( \langle \tilde{b}_{x1}; \ldots; \tilde{b}_{x128} \rangle \) as per equation (1).
A3. This step computes the block variation $\hat{\sigma}$ as per equation (2).
A4. This step computes the corresponding $\hat{r}$ as per step (A4).
A5. Denote by $f^r_x$ the recomputed value as per step (A4). Due to the chosen $\hat{SK}$ that was computed in step (D3) to ensure that a recomputed block feature matches the one embedded in the image, thus the recomputed $f^r_x$ will match the embedded $f^r$. 

B2. Since the recomputed $f^r_x$ equals the $f^r$ embedded in the watermarked image, the authentication check passes, and since the provided secret key $\hat{SK}$ has allowed this pass, the adversary is taken to be the rightful owner of the image. This attack requires an offline computation (step D3) with success probability $2^{-r}$; significantly higher than the probability of $2^{-k}$ in Claim 2.

The intuition exploited here is that the actual folded block feature $f^r_x$ embedded into the watermarkable pixel $p^x$ is only an $r$-bit ($2 \leq r \leq 4$) value. In the computation of $f^r$, the only unknown is $K$-bit $SK$ (stated as $k=168$ bits in [5]), thus it would appear that in order to compute the correct $f^r$ after modifying the unwatermarked pixels, requires to guess all $k=168$ bits of $SK$.

Nevertheless, due to the folding process used to produce $f^r$, this is not required. Instead, since $f^r$ is only $r$ bits ($2 \leq r \leq 4$), an attacker need only try a few values for the secret key to recompute $f^r$ with the tampered values of the unwatermarked pixels; and with a probability $2^{-r}$ he will compute the same $f^r$. Thus repeating only $2^{r}$ times (recall that this ranges from $2^{2}$ to $2^{4}$ times only) he will know the appropriate $\hat{SK}$ to use, even though it is a different $SK$ than originally used by the image owner.

3.3 Non-Binding Image Ownership

We highlight another security issue for the Chang et al. scheme. Ownership of the image is assumed based on the knowledge of the secret key $SK$ used to perform the watermark embedding process and later image authentication process. However, since $SK$ is secret and known only to a single person, there is nothing that binds that secret $SK$ to the identity of that person, and so he has no way to undeniably prove that he owns an image.

**Proposition 3 (Non-Binding Ownership)**. The Chang et al. scheme cannot provide binding ownership: an adversary can lay ownership claim without knowing the value of the owner’s secret key $SK$, therefore disproving Claim 2.

**Proof.** We illustrate with an attack description. Let the image watermarked by its owner be denoted as $I$.

E1. Perform steps (A1) through (A5) on the watermarked image $I$, where instead of the owner’s secret key $SK$, the adversary uses an arbitrarily chosen value $\hat{SK}$.

E2. Now perform step (A6) by inserting the computed block feature $f^r_x$ into the $r$ least significant bits of the watermarkable pixel $p^x$, overwriting the initially embedded $f^r$; to obtain the watermarked image $\hat{I}$. 


Consider an adversary who lays ownership claim on the image \( \tilde{I} \). He supplies his secret key \( \tilde{SK} \) and indeed, anyone performing the image authentication process would correctly be able to authenticate the image \( \tilde{I} \), so implying that the adversary rightfully owns the image, even though he does not know the value of the owner’s secret key \( SK \). This disproves Claim 2.

Indeed, it appears that in order for this scheme to really provide rightful ownership, that a trusted third party (TTP) is required to vouch that a particular secret key \( SK \) binds to an image. Yet, various watermarking schemes as early as 1998 already exist in literature [8] that achieve rightful ownership without requiring a TTP.

4 Concluding Remarks

We have demonstrated how to tamper with both the watermarked and unwatermarked pixels of an image that has been processed with the image authentication scheme of Chang et al. [5]. We also highlighted that the scheme does not provide unique binding of an image to an owner. Our results show that the scheme is not secure for its designed application of image authentication and image ownership. We suggest a thorough redesign by the authors before it can be used for practical applications, or alternatively to use other image authentication schemes in literature e.g. [1–3, 5, 6, 14, 25].

Acknowledgements

We thank the anonymous referees for suggesting to present the attack descriptions in a more formal way, and to highlight clearer the problems being exploited.

References