Pattern Recognition 41 (2008) 3493 – 3496

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Tampering with a watermarking-based image authentication scheme☆

Raphael C.-W. Phan*

Electronic and Electrical Engineering, Loughborough University, Loughborough LE11 3TU, UK

ARTICLE INFO

Article history:
Received 17 January 2008
Received in revised form 3 April 2008
Accepted 8 May 2008

PACS:
2.005

Keywords:
Encoding
Watermarking
Information hiding
Digital rights management
Image authentication
Watermarking
Tampering
Ownership

ABSTRACT

We analyse a recent image authentication scheme designed by Chang et al. [A watermarking-based image ownership and tampering authentication scheme, Pattern Recognition Lett. 27 (5) (2006) 439–446] whose first step is based on a watermarking scheme of Maniccam and Bourbakis [Lossless compression and information hiding in images, Pattern Recognition 37 (3) (2004) 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.

1. Introduction

An image authentication scheme allows to detect if any tampering has been performed on an image. Examples of watermarking-based [1–7,9–12,15–25] image authentication schemes are in Refs. [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 greyscale image I of $M \times N$ pixels is divided into $M/2 \times N/2$ overlapping blocks of size 3×3 pixels. The centre of each block is the watermarkable pixel p^x (where x is the block index) used for watermarking in each block.

(A2) For each watermarkable pixel p^x, the extracted block feature is actually the values of its eight neighbours (as in standard computer graphics terminology) denoted $\langle p_1^x, \ldots, p_8^x \rangle$; 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 3×3 block, formed by the dotted boxes and p^x in the middle.

☆ This research was done while the author was with the Security; Cryptography Lab (LASEC), EPFL, Switzerland.

* Tel.: +44 1509 227 080; fax: +44 1509 227 014.
E-mail address: r.phan@lboro.ac.uk.

0031-3203/$30.00 © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2008.05.009

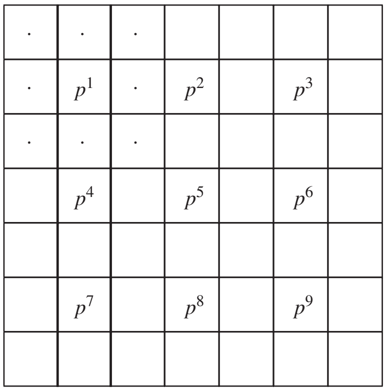

Fig. 1. Dividing the image into 3 × 3 blocks.

To be precise, the block feature is computed as

$$\langle b_1^x, \ldots, b_{128}^x \rangle = H(p_1^x \| \ldots \| p_8^x \| x \| ID \| SK), \tag{1}$$

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

(A3) The block variation σ is next computed as

$$\sigma = \sum_{i=1}^{8} (p_i^x - p_{i+1 \bmod 8}^x)^2. \tag{2}$$

(A4) r is then determined as follows:

$$r = \begin{cases} 2, & 0 \leqslant \sigma < 8, \\ 3, & 8 \leqslant \sigma < 16, \\ 4, & 16 \leqslant \sigma < 255. \end{cases}$$

(A5) The 128-bit block feature $\langle b_1^x, \ldots, b_{128}^x \rangle$ is then padded with zeroes to a length of 132 bits, and folded (compressed) into r bits as follows:

$$f^x = \bigoplus_{i=0}^{(132/r)-1} (b_{ir+1}^x \| \ldots \| b_{ir+r}^x), \tag{3}$$

where $\bigoplus$ denotes exclusive-or sum.

(A6) f^x is then inserted into the r least significant bits of the watermarkable pixel p^x to obtain the watermarked pixel $\hat{p}^x$. The image I with its watermarkable 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 Eq. (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)–(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*

$(\frac{1}{2})^r$; $2 \leqslant r \leqslant 4$.

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 × 3 pixels.

For instance, a 7 × 7 pixeled image will be divided into 3 × 3 (i.e. 9) blocks of 3 × 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 eight neighbours of p^1 are indicated by boxes with dots in them. Block 1 is a block of 3 × 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 $\hat{p}^x$, 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)–(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 Eq. (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 σ 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 σ, again the same value is computed for both the tampered and untampered watermarked pixels.

(A5) Denote by f'^x the recomputed value as per Eq. (3). Since this equation is a function of the block features $\langle b_1^x, \ldots, b_{128}^x \rangle$, hence the f'^x 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 p^x and embedded into the watermarkable pixel during the *watermark embedding* process.