Fragile Watermarking Based on Robust Hidden Information

ZHANG Xin-Peng(张新鹏), WANG Shuo-Zhong(王朔中)
School of Communication and Information Engineering, Shanghai University, Shanghai 200072, China

Abstract Block-wise fragile watermarks can be used to reveal maliciously tampered areas in multimedia products. However a forged content containing a cloned fragile watermark can be constructed by using a series of watermarked data. To defeat this type of counterfeit attack, a novel fragile watermarking technique is proposed in which different pseudo-random data are selected for different host products, and the generated fragile watermark is dependent upon the selected information. While inserting the fragile watermark, the pseudo-random information is also robustly embedded into the host data. Because of the difference between the selected information, different watermarked data cannot be used to forge illegal contents containing a valid fragile watermark.

Key words digital watermarking, information hiding, fragile, robust.

1 Introduction

For intellectual property rights (IPR) protection, imperceptible watermark can be embedded into multimedia products[1-2]. As digital codes containing information about the owner, origin, status, and/or destination of the data, robust watermarks are useful for ownership verification. In contrast, fragile watermarks are intended for checking integrity and authenticity of digital content.

Specifically, a copyright violator may try to replace portions of original content with fake information. It is therefore important to be able to precisely locate the particular areas that have been modified. In order to achieve this, the host data may be divided into small blocks and a fragile watermark embedded into each block. On the detection side, any data block is considered as “being modified” if the embedded mark in the block has been destroyed[3,4].

However, it is possible for the attacker to forge an arbitrary content with each block containing a cloned mark derived from a number of fragile-watermarked data samples[5]. Suppose that a sufficient amount of watermarked contents produced by an identical block-wise fragile watermarking system with a same secret key are available to the attacker. The fake content, into which the attacker is to embed his cloned fragile mark, is also divided into blocks. The attacker is able to find marked blocks from the available watermarked contents, which have least deviation with respect to the target blocks. Then, the original target blocks are replaced with the corresponding marked blocks so that a counterfeit content is constructed with each block containing the fragile watermark. In this way, the block-wise fragile watermarking system is defeated.

To resist this type of attacks, two methods have been proposed in the literature. Holliman and Memon suggested that the embedded fragile watermark be dependent not only on the covert block but also on its neighboring blocks[6]. Another scheme[8] derives the fragile watermark from an image ID and a block index.

Dependence of the watermark upon the surrounding blocks, however, may reduce accuracy of the location of tampered regions since any modification also destroys fragile marks in the neighborhood of a target block. Meanwhile, it is sometimes unpractical to obtain IDs of a protected product in watermark detection. In this work, a new scheme based on robust hidden information is proposed in order to combat the counterfeit attack. Here the fragile watermark depends
upon the protected covert block as well as some robust information purposely hidden in the host data, without involving any neighboring blocks and image ID.

2 Methodology

2.1 Generation and insertion of the robust information

An \( m \) bit binary sequence, \( R = \{ r_1, r_2, \ldots, r_m \} \), is randomly selected for the host media to be protected. Assume that the length of \( R \) is sufficiently large (for example, when \( m > 32 \)), and each host signal is assigned with a unique \( R \). Divide the host data into \( t \) sections, and insert the selected information \( R \) repeatedly into each section. The insertion can be based on addition of pseuderandom sequence\(^7\) or quantization index modulation\(^8\). The embedded information should be robust enough so that any possible attack that does not destroy the commercial value of the host media will not affect its extraction.

2.2 Embedding fragile watermark

In order to locate any maliciously replaced or modified region, the host data is divided into \( N \) block. The block size is usually quite small, for example, \( 8 \times 8 \). In each block, the embedded fragile watermark is a one-way hash output of the most significant bits of the host samples (pixel in case of image), the information of block position (the block index) and the hidden information. Replace the lowest significant bits of the host with the hash output. Thus, a fragile watermark is embedded with indiscernible distortion. Fig. 1 sketches the embedding procedure.

![Fig. 1 Sketch of watermark embedding procedure](image)

2.3 Fragile watermark detection

The \( t \) versions of robust hidden information are extracted from the \( t \) sections of the detected data. Although they may be different, each bit of the hidden sequence \( R \) can be restored according to the bit values in the majority of the versions. Only if the marked data has not been extensively and intensely tampered, can the robust information \( R \) be perfectly restored. In this case, the fragile watermark in each block can be computed using the most significant bits and the one-way hash function known to the detector. When the difference between the computed result and the least significant bits exists, the block is considered as "modified", otherwise "not modified".

In particular, if the marked data have been tampered extensively and intensely enough that the robust information \( R \) cannot be correctly restored, it is concluded that the entire media content has been "modified".

As described in the above, different robust information is selected for different media data, and fragile watermarks are dependent on block positions. If an attacker swaps sections of a marked product to forge an illegal version, the embedded fragile watermarks in the involved blocks will also be destroyed.

3 An Implementation for Still Image

In this section, an implementation of the proposed method for still images is described, and experimental results are presented.

Assume that the size of a black-and-white host image is \( 256 \times 256 \), which is segmented into 16 sections, each sized \( 64 \times 64 \). Randomly select 32-bit informations and repeatedly insert into each section using the following technique.

In each section, \( 8 \times 8 \) block-DCT is performed. Coefficients at position \( (2,2) \) of each DCT block are chosen to form a new data group sized \( 8 \times 8 \). The data in the formed group is then shuffled pseudo-randomly prior to a second-layer DCT in order to remove correlation between adjacent components so that the second layer DCT coefficients have a uniform expected energy, and provide a key to prevent unauthorized detection. Select 32 coefficients after the second-layer DCT and let each coefficient carry one bit of the robust information as expressed in the following equation.

\[
c'_i = \begin{cases} \text{round}\left(\frac{c_i}{\Delta}\right) \cdot \Delta, & \text{if embedded bit is } 1 \\ \text{round}\left(\frac{c_i + \Delta/2}{\Delta}\right) \cdot \Delta - \frac{\Delta}{2}, & \text{if embedded bit is } 0 \end{cases}
\]

(1)

In the experiment, \( \Delta \) is selected as 60 to achieve suffi-