Signal Processing 90 (2010) 3026–3036

Contents lists available at ScienceDirect

Signal Processing

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

Reversible fragile watermarking for locating tampered blocks in JPEG images

Xinpeng Zhang *, Shuozhong Wang, Zhenxing Qian, Guorui Feng

School of Communication and Information Engineering, Shanghai University, Shanghai 200072, P. R. China

ARTICLE INFO

Article history:
Received 23 December 2009
Received in revised form
29 April 2010
Accepted 30 April 2010
Available online 7 May 2010

Keywords:
Fragile watermarking
Reversible data-hiding
JPEG image
Image authentication

ABSTRACT

This paper proposes a novel fragile watermarking scheme for JPEG image authentication. The watermark is generated by folding the hash results of quantized coefficients, and each block is used to carry two watermark bits using a reversible data-hiding method. Because modification to the cover is small, the visual quality of watermarked image is satisfactory. On the receiver side, one may attempt to extract the watermark and recover the original content. By measuring mismatch between the watermark data extracted from the received image and derived from the recovered content, the blocks containing fake content can be located accurately, while the original information in the other blocks is retrieved without any error as long as the tampered area is not extensive.

1. Introduction

The purpose of fragile watermarking is to check integrity and authenticity of digital products, to locate the tampered areas and to recover the original contents [1,2]. Various fragile watermarking schemes have been developed for still images in uncompressed formats. In block-wise fragile watermarking, the host image is always divided into small blocks and the mark, e.g., a hash of the principal content of each block, is embedded into the block itself [3,4]. If the image has been changed, the image content and the watermark extracted from the tampered blocks do not match with each other, therefore the tampered blocks can be identified. In general, block-wise fragile watermarking methods are capable of detecting replacement of an extensive area. However, this type of techniques can only identify tampered blocks, but not the tampered pixels. Some pixel-wise fragile watermarking schemes have been proposed to resolve this problem, in which the watermark information derived from gray values of host pixels is embedded into the host pixels themselves [5,6]. So, tampered pixels can be identified from the absence of watermark information they carry. However, since some information derived from tampered pixel values may coincide with the watermark, localization of the tampered pixels is not complete, i.e., detection of the tampering pattern is inaccurate. In [7], a statistical mechanism is introduced into fragile watermarking, and two different distributions corresponding to tampered and original pixels are used to precisely locate the tampered pixels. In [8], the embedded watermark data are derived both from pixels and blocks, and a receiver can first identify the tampered blocks and then use the watermark hidden in the rest of the blocks to find the specific pattern of modification. Since it possesses advantages of both block-wise and pixel-wise techniques, the performance in locating tampered pixels is better than that of [7]. Furthermore, watermarking approaches that allow reconstruction of the original content in the tampered areas have been proposed [9,10]. In these methods, the main content in a region is embedded into another region of the image. After detecting the malicious

* Corresponding author.
E-mail addresses: xzhang@shu.edu.cn (X. Zhang), shuowang@shu.edu.cn (S. Wang), zxqian@shu.edu.cn (Z. Qian), grfeng@shu.edu.cn (G. Feng).

0165-1684/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.sigpro.2010.04.027

modification, the data extracted from reserved regions can be exploited to recover the principal content of tampered areas.

Fragile watermarking schemes can be integrated with reversible data hiding techniques. In reversible data hiding, some additional data are embedded into the cover signal in an invertible manner so that the original content can be perfectly restored after the hidden data are extracted. For example, the least significant digits of pixel values in an L-ary system can be losslessly compressed to provide a space for accommodating additional data [11]. Using the difference expansion (DE) algorithm [12], differences between two adjacent pixels are doubled so that a new LSB plane without carrying any information of the original is generated. The hidden message together with a compressed location map indicating the properties of pixel pairs, but not the host information itself, is embedded into the generated LSB plane. Since the compression rate of the location map is high, and almost every pixel pair can carry one bit, the DE algorithm can embed a fairly large amount of secret data into a host image. Moreover, various techniques have been introduced into the DE algorithm to improve the payload-distortion performance, including generalized integer transform [13], histogram shift [14], prediction of location map [15], and simplification of location map [16]. When a digital signature of the host content is embedded as a fragile watermark using a reversible data hiding technique, a receiver can detect any modification to the marked medium if the embedded watermark has been tampered, otherwise the original host data can be retrieved without error. Using a framework of reversible fragile watermarking [17], one can either locate the modified area from a tampered image or perfectly recover the original content from an authentic image. In [18], a tailor-made watermark is embedded into the host image using the DE technique. Although a malicious modification may destroy part of the embedded watermark, the tampered areas can be located and the watermark data extracted from the remaining regions can be used to restore the host image without any error.

JPEG is a widely used compression standard for transmitting and storing digital images. A number of fragile watermarking schemes have been developed for JPEG image authentication. Since the host data in the JPEG format does not provide sufficient space to accommodate the watermark, it is a challenge to exactly locate the tampered area when keeping a low distortion due to watermark embedding. For example, [19] embeds only one bit into each block of a host JPEG image. In an authentication procedure, the extracted watermark is compared with the original watermark, and a mismatch indicates the tampered blocks. Because some watermark bits extracted from the tampered blocks may coincide with the original ones, these tampered blocks cannot be detected. Using the method in [20], four watermark bits are embedded into each host block, and the embedded bits are dependent on the content of a block and its 8 neighboring blocks. Although coincidence in the tampered blocks is avoided, false alarm will occur in the neighborhood of the modified areas. In [21], all LSB of the quantized DCT coefficients in each block are replaced with the watermark data derived from a chaotic system. This way, detection of the tampered block is accurate, but distortion due to watermark embedding is high. In the above-mentioned fragile JPEG watermarking methods, the embedded watermark is not removable. That means, even though the tampered area is correctly located, the recipient can only obtain the watermarked content in reserved area, but not the original content. However, distortion introduced by watermark embedding, no matter how small it is, is unacceptable to some applications, e.g., military or medical images. So, a JPEG authentication scheme capable of exactly locating tampered blocks and recovering the original authentic content is desirable.

Reversible data-hiding techniques in JPEG images have been studied. In [22], LSB of quantized DCT coefficients corresponding to medium frequencies are losslessly compressed to provide a spare space to carry additional data. The method proposed in [23] expands the histogram of quantized DCT coefficients to produce some new histogram pairs, each of which containing an original position and an expansion position. Then, the original and expansion positions are used to represent the additional 0 and 1 respectively. In another approach, the data-hider exploits the zero-value DCT coefficients to carry data. In [24], a triplet consisting of one non-zero coefficient and two zero-value coefficients is used to accommodate one additional bit, and the first zero-value coefficient is changed according to the embedded bit. Alternatively, a high-frequency coefficient with an original zero value in each block is modified to embed several bits [25]. On the receiver side, after the embedded data are extracted, these modified coefficients are forced to zero to recover the original image. The method proposed in [26] attempts to embed the secret message in zero-value quantized DCT coefficients of medium-frequencies. Moreover, the number of nonzero coefficients that participate in the embedding process is limited so that distortion caused by data hiding is low.

This paper proposes a novel JPEG image authentication scheme, which combines fragile watermarking and reversible data-hiding technique. Having obtained the watermarked JPEG image, one can perfectly recover the original content. If some area in the watermarked image has been altered, the blocks containing fake information can be accurately located and the original content in the other blocks that are not damaged by the tampering can be retrieved as long as the tampered area is not too extensive. In this scheme, the watermark is generated by folding the hash results of quantized coefficients in image blocks as a short bit-sequence, and each block is employed to carry two watermark bits using a reversible data-hiding method. Because modification to the cover is small, the quality of watermarked image is satisfactory. On the receiver side, after extracting the watermark and recovering the original content, mismatch between the extracted watermark data and the recovered content is used to identify the blocks with incorrect restorations so that the tampered blocks are located. It can be assured that the recovered content in the rest of the image is exactly the same as the original unmarked host data.