A HIERARCHICAL IMAGE AUTHENTICATION WATERMARK WITH IMPROVED LOCALIZATION AND SECURITY

Mehmet U. Celik[a], *Gaurav Sharma*[b], *Eli Saber*[b], *A. Murat Tekalp*[a]

[a] Electrical and Computer Engineering Dept., University of Rochester, Rochester, NY, 14627-0126, USA
[b] Xerox Corporation, 800 Phillips Road, Webster, NY, 14580, USA
{celik,tekalp}@ece.rochester.edu, gsharma@ieee.org, eli.saber@usa.xerox.com

ABSTRACT

Several fragile watermarking schemes presented in the literature are either vulnerable to vector quantization (VQ) counterfeiting attacks or sacrifice localization accuracy to improve security. Using a hierarchical structure, we propose a method that thwarts the VQ attack while sustaining the superior localization properties of blockwise independent watermarking methods. In particular, we propose dividing the image into blocks in a multi-level hierarchy and calculating block signatures in this hierarchy. While signatures of small blocks on the lowest level of the hierarchy ensure superior accuracy of tamper localization, higher level block signatures provide increasing resistance to VQ attacks. At the top level, a signature calculated using the whole image completely thwarts the counterfeiting attack. Moreover, "sliding window" searches through the hierarchy enable the verification of untampered regions after an image has been cropped.

1. INTRODUCTION

The ease and extent of digital image manipulation highlights the need for authentication techniques in applications where verification of integrity and authenticity of the image is essential. Digital signatures are commonly used for authentication [1]. For image authentication, watermarks typically afford increased functionality by allowing localization of image manipulations and providing direct embedding of the watermark in the image data.

Authentication watermarks that can detect even small changes on the image are called *fragile* watermarks. A well known fragile watermark is Wong's scheme [2], which embeds a digital signature of most significant bits of a block of the image into least significant bits of the same block. Despite the elegance of the algorithm and cryptographic security of the digital signatures, its blockwise independence was exploited by Holliman and Memon with a counterfeiting attack [3]. Since the introduction of VQ codebook attack, a number of modifications for the existing algorithms have been proposed [3] [4]. Nonetheless, most of these methods, either fail to effectively address the problem or sacrifice tamper localization accuracy of the original methods[1].

In this paper, a new fragile watermarking algorithm based on the Wong's scheme [2] is proposed. Using a spatial hierarchy, the method thwarts the VQ codebook attack while sustaining the superior localization properties and the public key structure of the original algorithm.

[1] Fridrich [5] recently proposed an alternate elegant solution to the problem of localization with fragile watermarks in the presence of VQ attacks.

2. BACKGROUND

2.1. Wong's Scheme

In [2], Wong proposed a public key fragile watermarking algorithm. Wong's scheme works by partitioning the image into a number of blocks and inserting a digital signature for authentication on a block-wise basis. In the embedding process, for each block, a digital signature is computed after truncation of the least-significant bits (LSBs). The signature for each block is then embedded in the LSBs in conjunction with a logo image. At the receiving end, the digital signature of each block is recomputed after truncation of LSBs and the block is then authenticated by validating each block using the embedded information in the LSBs of the block. Unauthenticated blocks are assumed to have been manipulated, which provides the mechanism for localization of image manipulations.

2.2. Vector Quantization Counterfeiting Attack

Holliman and Memon [3] proposed a counterfeiting attack on blockwise independent watermarking schemes. The attacker approximates an image for which he wishes to create a forgery by using a collage of authentic blocks from watermarked images. Since the embedding and authentication processes are blockwise, the collage image is authenticated by the verification algorithm. Given a large enough database of watermarked images, the attacker can ensure that the counterfeit image has the same visual appearance as his original unwatermarked image. Wong's scheme is vulnerable to this attack as shown in [3].

2.3. Countermeasures

Several modifications of Wong's scheme have been proposed as countermeasures against the vector quantization counterfeiting attack.

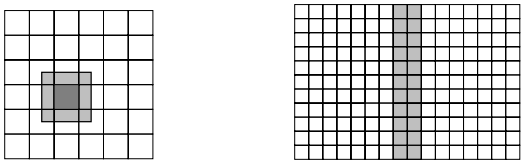
Increasing block dimensions:
Expected distortion and therefore the visual quality degradation caused by a vector quantization process depends on the size and the number of image blocks in a codebook. Smaller size blocks and larger codebook sizes yield better approximations. Therefore, the possibility of a reasonable forgery can be reduced by increasing the block dimensions used in the watermarking process. Larger blocks also decrease the number of authentic blocks that can be obtained from an image, further degrading the quality of the forgery by reducing codebook sizes. However, this countermeasure does not thwart the attack completely; if the set of watermarked images

available to the attacker is quite large, reasonable forgeries can still be produced. Moreover, using larger and larger blocks also impairs the tamper localization accuracy of the watermark.

Including block indices in the signature:
Wong's scheme may be slightly modified to include image indices in the signature computation step. Now, blocks of images at a different location cannot be used in the attack as a substitute. Yet it is possible to launch an attack given a large enough database of watermarked images.

Including image indices in the signature:
In [4], Wong and Memon suggest including also a unique image index in the signature. This method effectively prevents collating blocks from different images. However, the index value would also be necessary during verification. This constitutes an enormous burden in most of the practical applications. Considering such limitations, Wong and Memon suggests the extraction of the index from the image itself, e.g. as a hash of the whole image. Despite being a feasible alternative to index storage and management, this completely impairs the localization ability of the watermark. Manipulation in a single pixel of the image alters the calculated image index, which in turn invalidates signatures for all blocks.

Neighborhood dependent blocks:
An alternative method of eliminating the VQ counterfeiting attack is to eliminate the blockwise independence of the watermark. In particular, the signature embedded in a block may be calculated using a larger support, which overlaps the neighboring blocks (Fig. 1(a)). Using this scheme, a collage of individually watermarked blocks are no longer authenticated by the watermark extraction process because the neighboring blocks are not preserved. However, this modification results in some ambiguity in tamper localization. For instance, a possible result of the detection process is seen in Fig. 1(b), where shaded areas in the center are detected as non-authentic. This result may be obtained when: *i)* Center blocks of the image has been altered, *ii)* Parts of two different images are collated together. Therefore, it is not always possible to indicate the extent of the manipulation.



(a) Blockwise Dependence. (a) Tamper Localization.

Fig. 1. (a) Breaking blockwise independence. A larger support (shaded area) is used to calculate the signature which is embedded in (dark grey region). (b) Tamper localization in neighborhood dependent watermark. Unshaded areas are authenticated.

3. PROPOSED METHOD

We propose a hierarchical modification of Wong's scheme [2], which provides a graceful trade-off between security and tamper localization. In particular, we propose calculating signatures of the image blocks in a hierarchy.

3.1. Hierarchical Block-based Watermarking

Given an $N \times M$ image X, we first form a multi-level hierarchical block structure. Let us denote a block in this hierarchy by X_{ij}^{l}, where the indices ij represent the spatial position of the block and l is the level of the hierarchy to which the block belongs. The total number of levels in the hierarchy is further denoted by L.

On the lowest level, we partition the image X into $O \times P$ non-overlapping blocks $\{X_{00}^{L}, X_{01}^{L}, X_{10}^{L}, ..., X_{nm}^{L}\}$. At each successive level, the image is partitioned into blocks which in turn are composed of 2×2 blocks at the preceding level of the hierarchy,

for $l = L - 1$ to 2

$$\begin{bmatrix} X_{2i,2j}^{l+1} & \| & X_{2i,2j+1}^{l+1} \\ X_{2i+1,2j}^{l+1} & \| & X_{2i+1,2j+1}^{l+1} \end{bmatrix} = X_{ij}^{l}$$

Finally, top level of the hierarchy consists of only one block $X_{00}^{1} = X$. Note that we have larger blocks, in particular $2^{L-l}O \times 2^{L-l}P$, at upper levels of the hierarchy; no filtering or decimation is performed.

3.2. Watermark Insertion

The watermark insertion procedure consists of three main blocks as seen in Fig. 3: *i)* Formation of block hierarchy, *ii)* Computation of block signatures, and *iii)* Watermark insertion.

Upon formation of a proper hierarchy, for each block X_{ij}^{l}, a corresponding block $\tilde{X_{ij}^{l}}$ is formed by setting the least significant bit of each pixel to zero. Corresponding digital signatures are computed evaluating each pixel of the block $\tilde{X_{ij}^{l}}$ as a bit string. Only exception to the procedure is the top level block, where a *top* indicator is also included after the block. In general, this step consists of the calculation of hash H_{ij}^{l} of the block and public key encryption of the result.

for $l = 1$ to L,

$$H_{ij}^{l} = \mathcal{H}(\tilde{X_{ij}^{l}} \| [\text{“}top'']) \quad (1)$$

$$S_{ij}^{l} = Encrypt(H_{ij}^{l}, Key_{private}) \quad (2)$$

where ”top” is included only when $l = 1$.

Resulting signatures S_{ij}^{l} for each block are inserted into least significant bit-plane of the image. Since the blocks on different levels of the hierarchy share the same LSB plane, a partitioning algorithm that prevents any collision during insertion is required. A simple strategy is spreading high level signatures over a number of lower level blocks and inserting the accumulated payload at the lowest level of the hierarchy by LSB modification. Each lowest level block then carries a portion of upper level signatures, together with its independent signature. Thus, we proceed with partitioning the signature of each block into a number of smaller strings, where the exact number of such partitions is determined by the level of the block in the hierarchy.

$$S_{ij}^{l} = S_{ij}^{l}\{0,0\} \| S_{ij}^{l}\{0,1\} \| S_{ij}^{l}\{1,0\} \| ... \| S_{ij}^{l}\{\Lambda(l)-1, \Lambda(l)-1\} \quad (3)$$

where $\Lambda(l) = 2^{L-l}$. The number of lowest level blocks on which the signature is spread is then $\Lambda^{2}(l)$ Once atomic units are prepared, payload, P_{ij} of a block on the lowest level is formed by concatenating these units inherited from higher level blocks. This particular partitioning structure keeps the signature of the block at

each level localized inside the corresponding block. As a result, pixel manipulations outside a block does not effect the recovery of the signature and therefore the verification of the particular block.

Finally, LSB-plane of each block on the lowest level of the hierarchy is replaced by payload bits, P_{ij}. The watermarked image is a simple concatenation of these modified blocks. An illustration of the process is seen in Fig. 2.

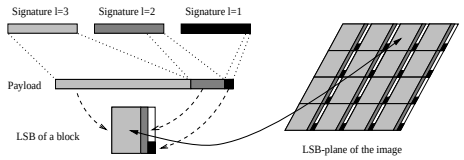


Fig. 2. Concatenation of signature blocks to form a payload (left) and spatial placement of resulting payload in LSB-plane of the image.

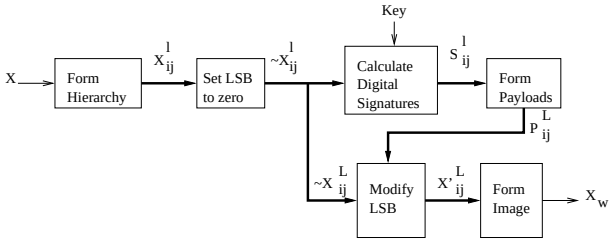


Fig. 3. Watermark insertion process for the proposed method

3.3. Watermark Verification

The watermark verification process consists of three basic steps analogous to the insertion procedure: *i)* Formation of block hierarchy, *ii)* Extraction of block signatures, and *iii)* Verification block signatures (Fig. 4).

Hierarchical block structure is formed as explained in Section 3.1. Payloads $\hat{P}_{ij}$ are extracted from the LSB-plane of each block at the lowest level. The partitioning algorithm used during insertion is reversed to recover all block signatures $\hat{S}^l_{ij}$.

For each block $\hat{X}^l_{ij}$, a quantized version $\hat{\tilde{X}}^l_{ij}$ is obtained by setting least significant bits of the pixels to zero. Reader should notice that $\tilde{X}^l_{ij}$ remains intact during watermark insertion; thus, unless the watermarked image is subsequently manipulated $\hat{\tilde{X}}^l_{ij}$ will be identical to $\tilde{X}^l_{ij}$.

At the last step, we verify the signature $\hat{S}^l_{ij}$. A block $\hat{X}^l_{ij}$ is deemed authentic if the signature $\hat{S}^l_{ij}$ verifies the quantized block $\hat{\tilde{X}}^l_{ij}$. A number of verification methods enabled by public key digital signature schemes may be utilized in this process. A general method consists of the following steps:

$$\hat{H}^l_{ij} = Decrypt(\hat{S}^l_{ij}, Key_{public}) \quad (4)$$

$$Verified = \begin{cases} True & \text{if } \hat{H}^l_{ij} = \mathcal{H}(\hat{\tilde{X}}^l_{ij}) \\ False & \text{otherwise} \end{cases} \quad (5)$$

As a result of signature verification step a hierarchical authenticity structure, an instance of which is seen in Fig. 8, is constructed. At the lowest level of the hierarchy, proposed method reduces to the original Wong's scheme with high tamper localization accuracy and susceptibility to a vector quantization counterfeiting attack. At each successive level, larger blocks yield lower resolution authentication maps with increasing resilience against counterfeiting attacks. Finally, the top level signature completely thwarts the counterfeiting attack.

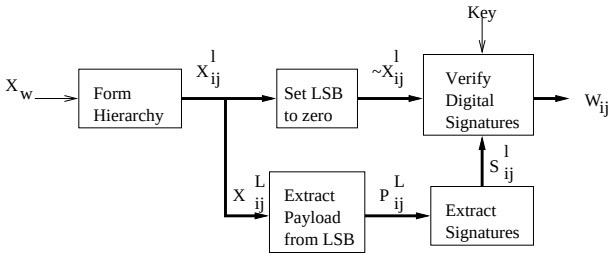


Fig. 4. Watermark verification process for the proposed method

3.4. Cropping Detection

Given an arbitrarily cropped image, it is desirable that a fragile watermarking scheme indicates the presence of cropping while still authenticating unaltered regions of the image. However, in most cases, watermark detection algorithm fails to verify even the authentic regions due to the loss of synchronization of block boundaries. For blockwise independent watermarking schemes, a "sliding-window" search can be utilized to regain synchronization with the block-boundaries as illustrated in Fig. 5. For the hierarchical scheme presented in this paper, a hierarchical search can be used to regain synchronization, detect the presence of cropping, and also authenticate untampered cropped regions. On the lowest level a sliding window search is performed in an $O \times P$ block. Once lowest level synchronization is regained, higher level searches are performed only using "sliding-block window" searches in 2×2 block neighborhoods.

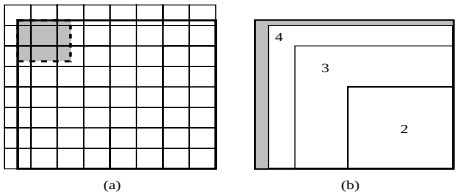


Fig. 5. Crop detection process. (a) A section of the watermarked image is cropped, seen in bold boundaries. Lowest level search is performed in the shaded area. (b) Corresponding watermark detection output. Shaded regions are not verified at any level. Complete blocks are verified at various levels of the hierarchy. Numbers show the lowest level of verification.

4. RESULTS

We implemented a private and a public key version of the proposed algorithm, differentiated by the digital signature scheme used. For the private key version, we use a 64 bit MAC (message authentication code) based on MD5 [1] algorithm as a digital signature, and for the public key version the 320 bit DSA [1] (digital signature algorithm) is employed.

An image is watermarked by the private key version of the algorithm yielding the watermarked image seen in Fig. 6. Watermarked image is then manipulated to yield the image seen in Fig. 7. In particular, the license plate of the car and the number on the door in the background are altered. In the third step, integrity and authenticity of the manipulated image is tested using the watermark detection algorithm, whose output is seen in Fig. 8. Numbers and shading indicate the lowest level signature verified, where darkest regions are not verified at any level. In Fig. 8, darkest regions obtained using the lowest level of the hierarchy contain the tampering in a small region. Higher level signature results confirm the response from the lowest level.

Fig. 6. Original watermarked image

Fig. 7. Manipulated image. License plate of the car and the number on the door in the background are altered.

Effectiveness of the proposed algorithm against a VQ attack has been demonstrated in the second test case. As in [3], a database of watermarked fingerprint images has been used for this attack. While the original unwatermarked image is seen in Fig. 9(a), counterfeit image is presented in Fig. 9(b). Output of our hierarchical method indicates that the attack is indeed successful at the lowest level of the hierarchy. Signatures of all blocks at this level are verified by our algorithm. Yet, on the higher levels of the hierarchy, block signatures cannot be verified. Evaluating the results as a whole we may confidently tell that counterfeiting attack is thwarted by the algorithm.

5. CONCLUSION AND DISCUSSION

In this paper we described a new hierarchical fragile watermarking scheme based on the public key watermark by Wong. The proposed method eliminates the vulnerabilities of the original scheme

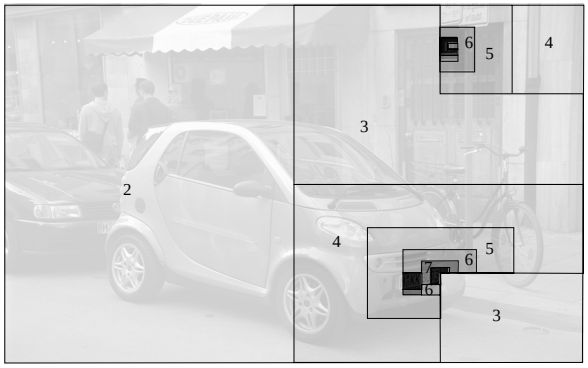


Fig. 8. Watermark detection output. Numbers and shading indicate the lowest level signature verified. Darkest regions are not verified at any level.

(a) Unwatermarked image (b) Counterfeit image

Fig. 9. Original and counterfeit images. Vector quantization attack uses 10×10 blocks from 19 watermarked images.

to VQ counterfeiting attack of Holliman and Memon [3]. As the attack effort is stepped up by using larger image blocks and larger image databases for the generation of counterfeit images, the hierarchical scheme gracefully sacrifices tamper localization accuracy while still detecting forgeries. This scheme also offers a significant advantage over most other watermarking schemes in that it allows for detection of cropping while still authenticating untampered cropped regions, albeit at a lower level of confidence.

6. REFERENCES

[1] A. Menezes, P van Oorchot, and S. Vanstone, *Handbook of Applied Cryptography*, CRC Press, Florida, USA, 1997.

[2] P.W. Wong, "A public key watermark for image verification and authentication," in *Proceedings of IEEE International Conference on Image Processing*, Chicago, USA, October 4-7, 1998, pp. 425–429.

[3] M. Holliman and N. Memon, "Counterfeiting attacks on oblivious block-wise independent invisible watermarking schemes," *IEEE Transactions on Image Processing*, vol. 9, no. 3, pp. 432–441, March 2000.

[4] P.W. Wong and N. Memon, "Secret and public key authentication watermarking schemes that resist vector quantization attack," *Proceedings of SPIE Security and Watermarking of Multimedia Contents II*, vol. 3971, no. 40, Jan 2000.

[5] J. Fridrich, Private communication, Feb 2001.