Hierarchical Watermarking for Secure Image Authentication With Localization

Mehmet Utku Celik, *Student Member, IEEE*, Gaurav Sharma, *Senior Member, IEEE*, Eli Saber, *Senior Member, IEEE*, and Ahmet Murat Tekalp, *Senior Member, IEEE*

***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 multilevel 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. We provide experimental results to demonstrate the effectiveness of our method.**

***Index Terms*—Authentication, fragile watermark, tamper localization, vector quantization attack.**

I. Introduction

TRADITIONALLY, due to the limited processing abilities in analog media, malicious manipulation of images has been a tedious task with only low quality results being realized without prohibitively expensive professional equipment. However, digital images, unlike their analog counterparts, can be easily manipulated using a variety of sophisticated image processing tools that are readily available as commercial packages. The ease and extent of such manipulations emphasize the need for image authentication techniques in applications where verification of integrity and authenticity of the image content is essential. Potential security loopholes of shared information networks, e.g., Internet, on which images are commonly posted and distributed further underscore this need.

Multimedia integrity and authenticity can be guaranteed through the use of *digital signatures* and/or watermarks. A *digital signature* is a data string which associates a message (in digital form) with some originating entity [1]. Digital signatures and their properties have been well studied in cryptography, and a number of algorithms, such as RSA and DSA, are extensively deployed in various authentication applications [1]. Digital watermarking (see [2]–[5]) may be utilized in general to verify authenticity and integrity of multimedia content. The use of watermarks instead of digital signatures typically affords additional functionality by exploiting inherent properties of image content. Examples of such advantages are the capability for localization of manipulations made to the image and the direct embedding of the watermark in the image data. It is worth mentioning that, both digital signatures and authentication watermarks are useful only for establishing the source of the image and detecting manipulations occurring after the signature/watermark has been inserted. However, neither technique by itself is capable of certifying that an image represents an original unaltered scene, unless supported by additional mechanisms [6].

Authentication watermarks can be classified as either *fragile* or *semi-fragile*. Fragile watermarks, as the name implies, are designed to identify any alteration of the pixel values. Semi-fragile watermarks, on the other hand, try to differentiate between content-preserving (nonmalicious) processes, e.g., compression, and malicious manipulations, e.g., removal of objects from a scene. Watermarks in this class are designed to withstand content-preserving operations, while detecting any malicious manipulations. Various algorithms have been proposed for fragile watermarking [7]–[10] and semi-fragile watermarking [11]–[13]. Though semi-fragile watermarks can provide extended functionality, in this paper, we will restrict our attention to fragile watermarks for which the issues of tamper localization and manipulation detection are well defined.

A general block diagram representing most fragile watermarking schemes is shown in Fig. 1, where the watermark embedding and extraction processes utilize cryptographic keys. Fragile watermarks are classified as *public key* and *private key* methods. Private key watermarks are symmetric key systems which use the same key for watermark embedding and extraction. The key is known only to the watermark embedder and, therefore, the verification of the authenticity and integrity of the image can also be done only by the watermarker alone. Public key watermarks, on the other hand, are asymmetric key systems that utilize a secret private key for watermark embedding and a corresponding publicly available key for the extraction. Public availability of the extraction key enables public detection of the watermark and thereby verification of authenticity and integrity of the image and tamper localization, which is typically desired in most fragile watermarking applications.

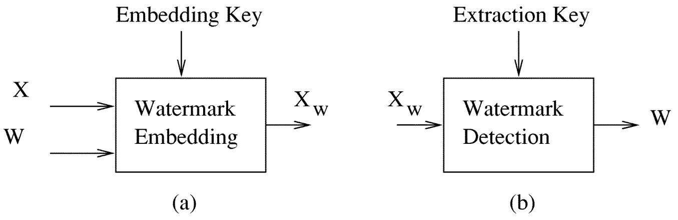
Manuscript received May 22, 2001; revised February 17, 2002. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Christine Guillemot.

M. U. Celik is with the Electrical and Computer Engineering Department, University of Rochester, Rochester, NY 14627 USA (e-mail: celik@ece.rochester.edu).

G. Sharma and E. Saber are with Xerox Corporation, Webster, NY 14580 USA (e-mail: g.sharma@ieee.org; eli.saber@usa.xerox.com).

A. M. Tekalp is with the Electrical and Computer Engineering Department, University of Rochester, Rochester, NY 14627 USA and also with the College of Engineering, Koc University, Istanbul, Turkey (e-mail: tekalp@ece.rochester.edu).

Publisher Item Identifier S 1057-7149(02)04791-7.



(a) (b)

Fig. 1. Fragile watermarking: (a) embedding and (b) detection. Authenticity of the image X is protected by embedding a watermark pattern W. Manipulations on the watermarked image X_w is observed through the changes on the detected watermark.

A well-known algorithm among public key fragile watermarks is Wong's scheme [8], which embeds a digital signature of the most significant bits of a block of the image into the 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 [14]. The attacker constructs a vector quantization codebook using blocks from a set of watermarked images. The image to be counterfeited is then approximated using this codebook. Since each block is authenticated by itself, the counterfeit image appears authentic to the watermarking algorithm. Since the introduction of VQ codebook attack, a number of modifications for the existing algorithms have been proposed [9], [14], [15]. 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, we propose a new fragile watermarking algorithm based on the Wong's scheme [8]. Using a special hierarchical structure, our method thwarts the VQ codebook attack while sustaining the superior localization properties and the public key structure of the original algorithm. The image to be watermarked is divided into blocks in a multilevel hierarchy. At the lowest level of the hierarchy, the image is partitioned into a set of elementary blocks composed of groupings of image pixels. At each successive level, the image is partitioned into blocks which in turn are composed of blocks at the preceding level of the hierarchy. At each level of the hierarchy, a digital signature (or cryptographic hash function) for each block is calculated using the seven most significant bit-plane values of all pixels within the block. The resulting signature is incorporated into the LSBs of selected pixels within the block. The selection of pixels for the embedding of signatures corresponding to a block at a given level of the hierarchy is done using a partition of the LSBs of each elementary block according to the chosen multilevel hierarchy. While independent block signatures localize manipulations at elementary block level, higher level signatures provide increasing resistance to VQ codebook attacks, gracefully trading-off accuracy of localization for greater security.

An alternative approach to the proposed multilevel (pyramid-like) hierarchical scheme would have been to embed the watermark in the wavelet transform domain (similar to the semi-fragile method of Kundur *et al.* [11]). However, a number of technical and implementation issues make it impractical to develop a fragile multiresolution watermarking scheme in the wavelet domain. First, in some cases, random perturbations of the LSB of the high frequency band coefficients may result in pixel values that are outside the dynamic range of the original image, which makes recovery of the watermark impossible. Moreover, data embedding in the LSB of the high frequency bands of the wavelet representation may not correspond to LSB modifications in the image domain, and the amount of distortion, hence visibility of the watermark, may increase by the number of stages that are modified in the wavelet transform.

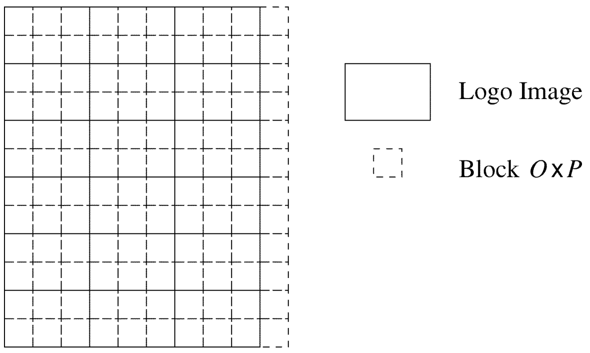


Fig. 2. Tiling of logo image in Wong's scheme.

The rest of the paper is organized as follows: In Section II, we discuss Wong's original scheme, vector quantization counterfeiting attack, and proposed countermeasures against this attack. Our hierarchical watermarking method is proposed in Section III. We present our experimental results and an analysis of the algorithm in Section IV. Conclusions are drawn in Section V.

II. Background

A. Authentication Watermark by Wong

Wong's scheme [8] is a block-based watermarking technique. In this scheme, given an $M \times N$ image X, a binary watermark image W of the same size is initialized. In practice, this step is usually achieved by tiling the original image with a smaller logo image, as illustrated in Fig. 2.

The original image X is partitioned into $O \times P$ pixel blocks, $\{X_1, X_2, \ldots\}$; where X_r denotes such blocks. Likewise, the watermark image is partitioned into blocks, W_r. For each block X_r, a corresponding block $\tilde{X}_r$ is formed by setting the least significant bit of each pixel to zero. A cryptographic hash, e.g., MD5 or SHA [1], of transformed block $\tilde{X}_r$ and image dimensions is computed

$$H_r = \mathcal{H}\left(M,\, N,\, \tilde{X}_r\right). \tag{1}$$

The signature of a block is formed by XORing the computed hash with the watermark pattern and encrypting the result with a public key encryption algorithm

$$S_r = Encrypt(H_r \oplus W_r,\, Key_{private}) \tag{2}$$

where $\oplus$ denotes the bitwise XOR operator. Finally, the signature S_r is inserted in X_r as the least significant bits of the

[1] During the course of writing this paper, the authors became aware of recent independent work by Fridrich [16] which provides an alternate elegant solution to the problem of localization with fragile watermarks in the presence of VQ attacks.

block. Note that the application of this procedure independently on each block produces the watermarked image.

During watermark verification similar steps are followed. First the candidate image $\hat{X}$ is partitioned into blocks $\hat{X}_r$. Signature $\hat{S}_r$ is read from the least significant bits of each block, $\hat{X}_r$. $\tilde{\hat{X}}_r$s are formed by setting LSBs to zero and $\hat{H}_r$s are calculated using image sizes and $\tilde{\hat{X}}_r$s. Finally, watermark image blocks are recovered by XORing the hash values with decrypted signatures from each block

$$\hat{W}_r = Decrypt\left(\hat{S}_r,\, Key_{public}\right) \oplus \hat{H}_r. \tag{3}$$

Any changes in the pixel values of a block alter either the decrypted signature or output of the hash function. Thus, any manipulation on the image is detected by the change in the corresponding region of the binary watermark image.

B. Vector Quantization Counterfeiting Attack

Holliman and Memon [14] 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 collage image has the same visual appearance as his original unwatermarked image.

In order to explain this attack, let us define *blockwise independence* and K*-equivalence* first. Note that, the terminology and notation adopted here is similar to that utilized in [14]. A watermarking technique is block-based if it partitions an image X into nonoverlapping blocks $\{X_1,\, X_2,\, \ldots,\, X_n\}$ and inserts a watermark W_i in block X_i using the key K_i. A block-based technique is *blockwise independent* if each watermarked block $\acute{X}_i$ depends only on the original block X_i, the watermark W_i and the insertion key K_i.

Thus, watermark embedding function $\mathcal{E}_K()$ and the detection function $\mathcal{D}_K()$, which operates on an potentially altered image $\hat{X}$, can be represented by

$$\begin{aligned} \acute{X} &= \mathcal{E}_K(X,\, W) \\ &= \mathcal{E}_{K_1}(X_1,\, W_1) \| \mathcal{E}_{K_2}(X_2,\, W_2) \| \cdots \| \mathcal{E}_{K_n}(X_n,\, W_n) \end{aligned} \tag{4}$$

$$\begin{aligned} \hat{W} &= \mathcal{D}_K\left(\hat{X}\right) = \mathcal{D}_{K_1}\left(\hat{X}_1\right) \left\| \mathcal{D}_{K_2}\left(\hat{X}_2\right) \right\| \cdots \left\| \mathcal{D}_{K_n}\left(\hat{X}_n\right) \right. \\ &= \hat{W}_1 \left\| \hat{W}_2 \right\| \cdots \left\| \hat{W}_n \right. \end{aligned} \tag{5}$$

where $K_1,\, K_2,\, \ldots,\, K_n$ are embedding keys which may be derived from a single key K, $W = W_1 \| W_2 \| \cdots \| W_n$ is the watermark pattern and $\|$ denotes concatenation.

Furthermore, two image blocks X_i and X_j are said to be K-*equivalent* if a given key K extracts the same watermark from both of them. That is

$$\mathcal{D}_K(X_i) = \mathcal{D}_K(X_j) = W. \tag{6}$$

Hence, given a key K, any blockwise independent watermarking method partitions the set of all image blocks into equivalence classes $\{C_1,\, C_2,\, \ldots,\, C_m\}$, where m is the number of different possible watermark signals. The application of a watermark detection process to any block from a given equivalence class C_i results in the same watermark being recovered with the key K.

The attack exploits the K-equivalence property of blockwise independent watermarking schemes. Suppose the attacker wants to counterfeit an image Y, and has access to one or more images, say $\acute{X}$, watermarked by a watermark image W and a key K. The attacker can construct a counterfeit image $\acute{Y}$ which is sufficiently similar to Y as follows:

```
Let $Y = Y_1 \| Y_2 \| \cdots \| Y_n$ and
$X = X_1 \| X_2 \| \cdots \| X_n$ be a watermarked image of
  the same size.
for $i = 1$ to $n$
  Identify equivalence class $k$ of $X_i$.
  Find an approximation $\acute{Y}_i \in C_k$ to $Y_i$ such
  that $\acute{Y}_i \simeq Y_i$.
  Replace $Y_i$ by $\acute{Y}_i$.
Construct $\acute{Y} = \acute{Y}_1 \| \acute{Y}_2 \| \cdots \| \acute{Y}_n$.
```

In most cases, only partial knowledge of the watermark image W, e.g., the tiling of logo images in Wong's scheme, is sufficient to classify the blocks of the watermarked images into the different equivalence classes C_k. Thus the watermarked images $\acute{X}$ can be used to populate subsets of the equivalence classes, which can then be used in the approximation process described above. These subsets may be viewed as vector-quantization (VQ) codebooks, with the codebook corresponding to an equivalence class C_k composed of the blocks from the watermarked images $\acute{X}$ that are in C_k. The process of approximating Y_i by $\acute{Y}_i$ such that $\acute{Y}_i \simeq Y_i$ and $\acute{Y}_i \in C_k$ can be interpreted as vector quantization of Y_i using the codebook corresponding to C_k.

VQ attack on Wong's scheme is performed similarly. Partitioning of the binary watermark image composed of tiles effectively partitions the logo image into blocks (Fig. 2). Thus, each distinct block of the logo represents an equivalence class in the attack. A vector quantization codebook is constructed for each such class by properly assigning blocks of the watermarked images to an equivalence class. Resulting codebooks can be used to carry out steps of the attack, which are explained above. An example demonstrating the success of such an attack is seen in Fig. 15.

C. Countermeasures Against Counterfeiting Attack on Wong's Scheme

In this section, we will elaborate on a number of modifications on Wong's scheme which 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 two key factors: the size and the number of image blocks in a codebook. Smaller size blocks can be approximated more accurately given a fixed size codebook. Similarly, better approximations can be obtained as the number of blocks in the codebook increases. 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 one fixed-size image, further degrading the quality of the forgery by reducing codebook sizes.

This countermeasure, however, 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

$$H_r = \mathcal{H}\left(M, N, \tilde{X}_r, r\right) \tag{7}$$

$$S_r = Encrypt(H_r \oplus W_r, Key_{private}). \tag{8}$$

This effectively increases the number of equivalence classes. Now, blocks from different locations belong to separate equivalence classes. The codebook for each class is limited to the blocks with the same index, yet it is possible to launch an attack given a large enough database of watermarked images.

- *Including Image Indices in the Signature*

  In [15], Wong and Memon suggest including also a unique image index in the signature. This further increases the number of equivalence classes and restricts the codebook construction domain. Using sufficiently large index values, counterfeiting attack can be practically eliminated

$$H_r = \mathcal{H}\left(M, N, \tilde{X}_r, r, ID\right) \tag{9}$$

$$S_r = Encrypt(H_r \oplus W_r, Key_{private}). \tag{10}$$

However, it should be noted that such an index would also be necessary during verification. While managing such individual indices for all images in a database may be possible for some applications, in most of the practical applications this constitutes an enormous burden. 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 results in different hash H_r values for all blocks. Effectively, smallest change in the image distorts all of the extracted watermark.

An alternative approach to overcome the challenge of image index storage has been proposed by Fridrich *et al.* [9]. In their method, an image index is embedded within the image in multiple positions. In case of a manipulation, the multiple copies increase the chances of extracting the correct image index during watermark detection. Nevertheless, this does not guarantee correct index extraction in all cases. A manipulation that is spatially restricted within regions carrying the index values may lead to extraction of corrupt index values. Since these values are used during verification of all image regions, such a manipulation may cause the entire image to be invalidated even though the manipulation altered only a small part of the image and the major portion of the image is unaltered from the authentic version. The embedding of multiple copies also either increases the embedding distortion or reduces the localization ability of the watermark by consuming capacity that is normally utilized by the digital signatures.

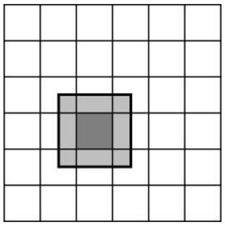


Fig. 3. Breaking blockwise independence. A larger support $\mathcal{X}_r$ (shaded area) is used to calculate the signature which is embedded in X_r (dark grey region).

- *Breaking Block-Wise Independence: Neighborhood Dependent Blocks*

  Increasing the number of equivalence classes requires larger databases of watermarked images for the construction of good VQ codebooks, thereby making the VQ counterfeiting attack impractical in most cases. 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 X_r may be calculated using a larger support $\mathcal{X}_r$, which overlaps the neighboring blocks (Fig. 3). This technique is very similar to block chaining modes used in block encryption techniques, e.g., CBC mode in DES [1]. Using this scheme, a collage of individually watermarked blocks of an image is no longer authenticated by the watermarking extraction process because the larger support covering the neighboring blocks is not preserved.

  Though the use of neighborhood dependent blocks eliminates the possibility of a forgery going undetected, the method results in some ambiguity in tamper localization. For instance, Fig. 4 shows a possible result of the detection process using the proposed modification. Only shaded areas in the center are detected as nonauthentic, which may have resulted from two different manipulations:

  — Center blocks of the image has been altered
  — Parts of two different images are collated together.

  Therefore, in this case, it is not possible to indicate the extent of the manipulation.

III. Proposed Method

We propose a hierarchical modification of Wong's scheme, which provides a graceful trade-off between security and

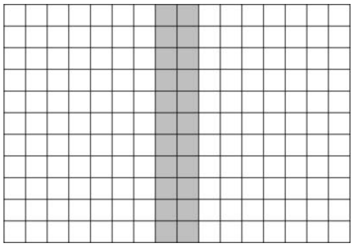


Fig. 4. Tamper localization in neighborhood dependent watermark. Unshaded areas are authenticated.

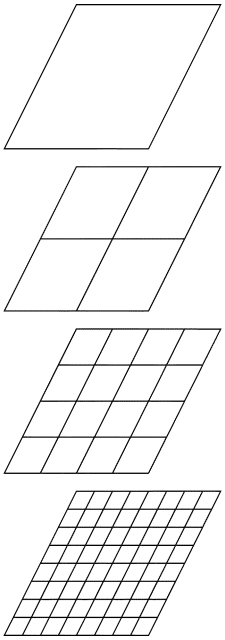


Fig. 5. Partitioning of an image and the resulting four level hierarchical block structure.

tamper localization. In particular, we propose calculating signatures of the image blocks in a hierarchy. We first describe the proposed hierarchical watermarking scheme. The watermark insertion and extraction processes and cropping detection are presented thereafter.

A. Hierarchical Block-Based Watermarking

A *hierarchical block-based watermarking* technique inserts and extracts a watermark in a multilevel hierarchy. Partitioning of the image into nonoverlapping blocks constitutes the lowest level of the hierarchy (Fig. 5). Successive levels of the hierarchy are formed by combining distinct groups of blocks at a preceding level of the hierarchy. In general, the number of blocks from a lower level of the hierarchy that are combined to form a block at the next level of the hierarchy may be arbitrarily chosen, however, in order to keep the notation and the description simpler, we assume for the rest of this paper that the region of 2 × 2 blocks at a given level of the hierarchy are combined to create a block at the next level of the hierarchy.

Given an $N \times M$ image X, we first form a multilevel 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 $O \times P$ nonoverlapping blocks $\{X_{00}^{L}, X_{01}^{L}, X_{10}^{L}, \ldots, X_{nm}^{L}\}$. 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. That is,

$$\text{for } l = L-1 \text{ to } 2$$
$$\begin{bmatrix} X_{2i,2j}^{l+1} & \| & X_{2i,2j+1}^{l+1} \\ X_{2i+1,2j}^{l+1} & \| & X_{2i+1,2j+1}^{l+1} \end{bmatrix} = X_{ij}^{l}.$$

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.

B. Watermark Insertion

The watermark insertion procedure consists of three main blocks as seen in Fig. 6: i) Formation of block hierarchy as described above, 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

$$\text{for } l = 1 \text{ to } L,$$
$$H_{ij}^{l} = \mathcal{H}\left(\tilde{X}_{ij}^{l} \| [top]\right) \tag{11}$$
$$S_{ij}^{l} = Encrypt\left(H_{ij}^{l}, Key_{private}\right) \tag{12}$$

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. For instance, in a hierarchy of three levels and a digital signature of $\mathcal{S}$ bits, lowest level block carries $(21/16)\mathcal{S}$ bits which consists of $\mathcal{S}$, $\mathcal{S}/4$, $\mathcal{S}/16$ bits corresponding to the entire signature of itself, one fourth of the upper level block’s and one sixteenth of the top level block’s, respectively. Thus, we proceed with partitioning the signature of each block into a number of smaller strings, where the exact number of

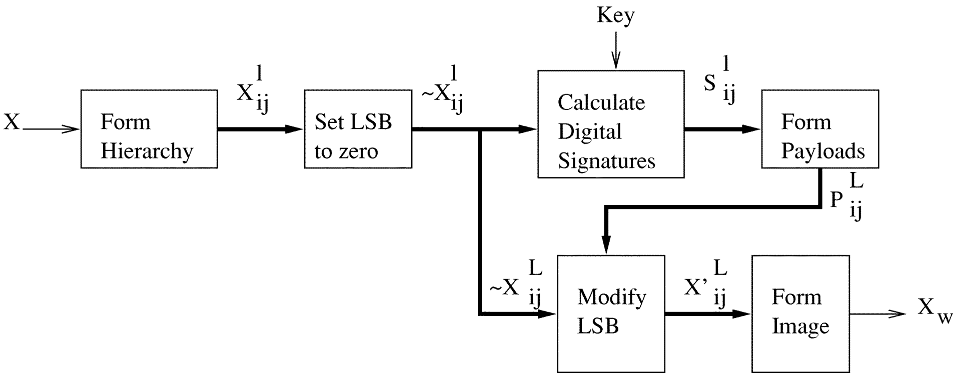


Fig. 6. Watermark insertion process for the proposed method.

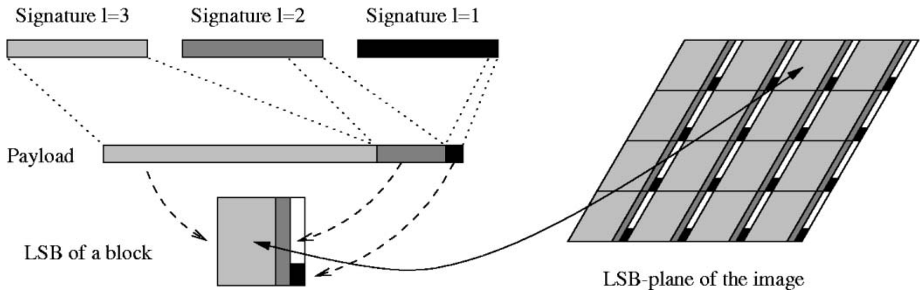


Fig. 7. Concatenation of signature blocks to form a payload (left) and spatial placement of resulting payload in LSB-plane of the image.

such partitions is determined by the level of the block in the hierarchy

$$S^l_{ij} = S^l_{ij}\{0, 0\} \| S^l_{ij}\{0, 1\} \| S^l_{ij}\{1, 0\} \| \cdots \| S^l_{ij}\{\Lambda(l) - 1, \Lambda(l) - 1\} \tag{13}$$

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 of a block on the lowest level is formed by concatenating these units inherited from higher level blocks

$$P_{ij} = S^L_{ij} \| S_{Q_{L-1}(i),\, Q_{L-1}(j)}\{i - Q_{L-1}(i),\, j - Q_{L-1}(j)\} \| \cdots \| S_{Q_1(i),\, Q_1(j)}\{i - Q_1(i),\, j - Q_1(j)\} \tag{14}$$

$$Q_k(x) = \left\lfloor \frac{x}{2^{L-k}} \right\rfloor. \tag{15}$$

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 do 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. Let $\acute{X}^L_{ij}$ denote modified blocks. The watermarked image $\acute{X}$ is a simple concatenation of these blocks. An illustration of the process is seen in Fig. 7

$$\acute{X} = \begin{bmatrix} \acute{X}^L_{00} & \acute{X}^L_{01} & \cdots & \acute{X}^L_{0n} \\ \acute{X}^L_{10} & \acute{X}^L_{11} & \cdots & \acute{X}^L_{1n} \\ \cdots & \cdots & \cdots & \cdots \\ \acute{X}^L_{m0} & \acute{X}^L_{m1} & \cdots & \acute{X}^L_{mn} \end{bmatrix}. \tag{16}$$

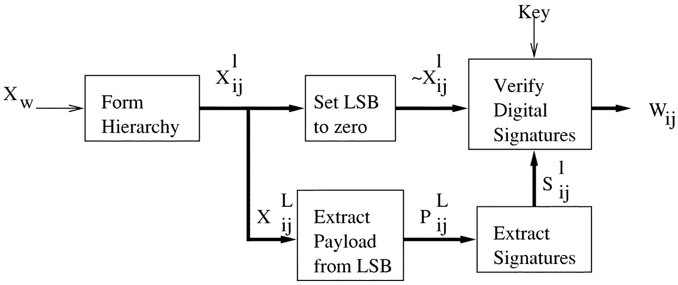


Fig. 8. Watermark verification process for the proposed method.

C. 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. 8).

Hierarchical block structure is formed as explained in Section III-A. 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 $\tilde{\hat{X}}^l_{ij}$ is obtained by setting least significant bits of the pixels to zero. The reader will notice that $\tilde{X}^l_{ij}$ remains intact during watermark insertion; thus, unless the watermarked image is subsequently manipulated $\tilde{\hat{X}}^l_{ij}$ 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 $\tilde{\hat{X}}^l_{ij}$. A number of verification methods enabled by public

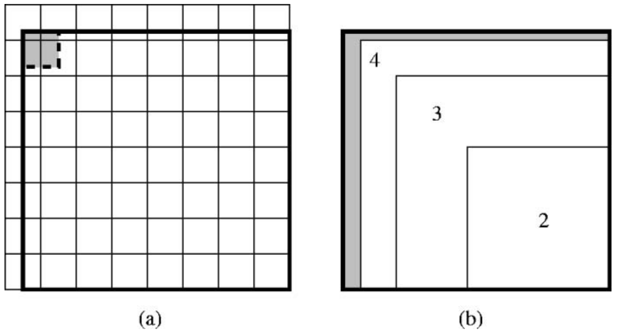


Fig. 9. Crop detection process. (a) A section of the watermarked image is cropped, seen in bold boundaries. Lowest level search is performed in the shaded area. Once the block boundary is synchronized at the lowest level, higher level searches are done in 2 × 2 neighborhoods. (b) Corresponding watermark 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.

key digital signature schemes may be utilized in this process. A general method consists of the following steps:

$$\hat{H}^l_{ij} = Decrypt\left(\hat{S}^l_{ij},\, Key_{public}\right) \tag{17}$$

$$Verified = \begin{cases} True, & \text{if } \hat{H}^l_{ij} = \mathcal{H}\left(\tilde{\hat{X}}^l_{ij}\right) \\ False, & \text{otherwise.} \end{cases} \tag{18}$$

As a result of the signature verification step, a hierarchical authenticity structure, an instance of which is seen in Fig. 12, is constructed. At the lowest level of the hierarchy, the 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. The top level signature does not enable any tamper localization, however it completely thwarts the possibility of a counterfeiting attack. A secure image authentication scheme with good localization properties is achieved when the results of the signature verification step are evaluated altogether.

D. Cropping Detection

Cropping is one of the simplest image manipulations that may be performed, wherein a smaller rectangular region of a larger image is extracted and the remaining portions discarded. 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. The detection of cropping has however received only limited attention in fragile watermarking so far. In most cases, the watermark detection algorithm will fail 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. 9. 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.

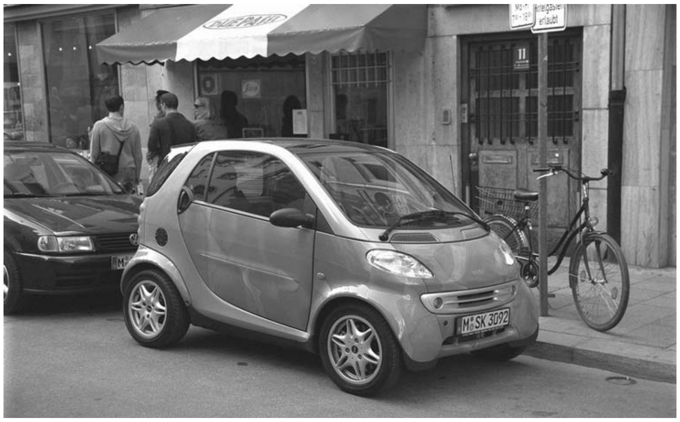


Fig. 10. Original watermarked image.

A shortcoming of the proposed scheme can be observed when only a quarter of the image has been cropped. In this case, all the block signatures matches the original signatures at all levels. However, the slight modification of the signature formation for the top level (a "top" indicator is included) eliminates the possibility of cropping going undetected. The proposed method properly differentiates a part of the original image from the whole image.

IV. Results and Analysis

We proposed a novel hierarchical authentication watermark in the preceding section. Now, we will demonstrate the effectiveness of our method with experimental results and discuss the performance of the algorithm.

A. Experimental 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 algorithm as a digital signature, and for the public key version the 320 bit DSA (digital signature algorithm) is employed. Details of the cryptographic functions mentioned here may be found in common cryptography texts such as [1].

In the first test case, we demonstrate the localization and tamper detection ability of our algorithm. An image is watermarked by the private key version of the algorithm yielding the watermarked image seen in Fig. 10. During embedding, the 832 × 512 gray-scale image has been decomposed into a seven-level hierarchy with 13 × 8 blocks at the lowest level. Note that LSBs of a lowest level block may be modified to carry a payload of $13 \times 8 = 104$ bits, which is sufficient to accommodate the $4/3 \times 64 \simeq 86$ bits required by the algorithm. At each successive level, block dimensions are doubled, until the top level consists of the whole image. Watermarked image is then manipulated using image processing software to yield the image seen in Fig. 11. In particular, the license plate of the car and the number on the door in the background are altered. Magnified versions of the manipulated regions are presented in

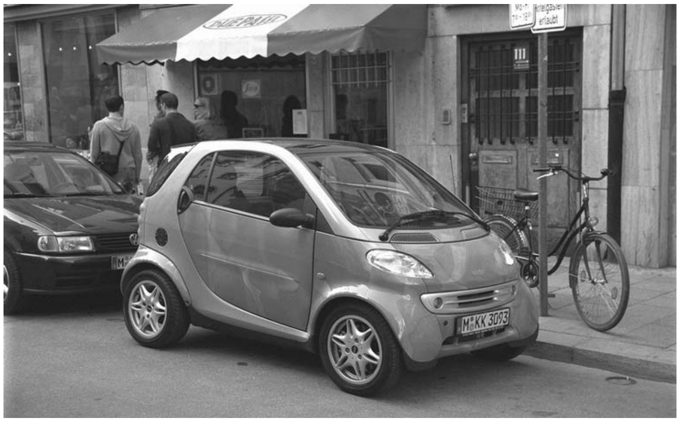


Fig. 11. Manipulated image. License plate of the car and the number on the door in the background are altered.

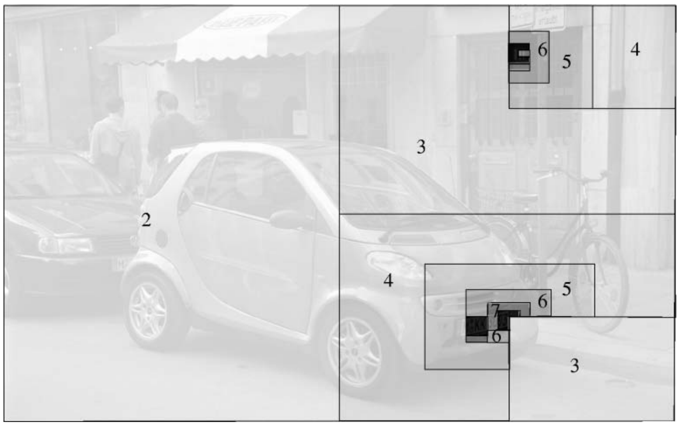


Fig. 12. Watermark detection output. Numbers and shading indicate the lowest level signature verified. Darkest regions are not verified at any level.

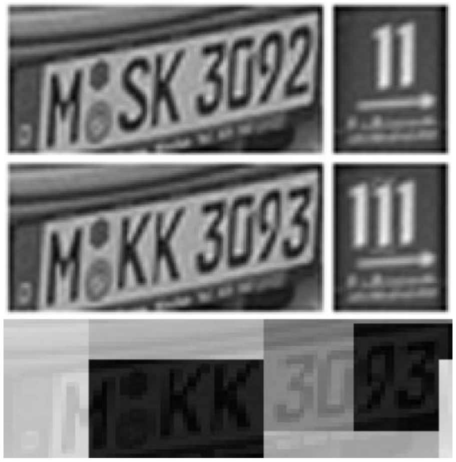


Fig. 13. License plate of the car and number on the door: Original (top), Manipulated (center), Detection output (bottom).

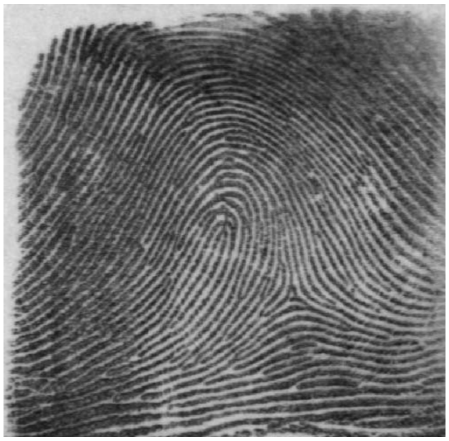


Fig. 14. Original unwatermarked fingerprint image.

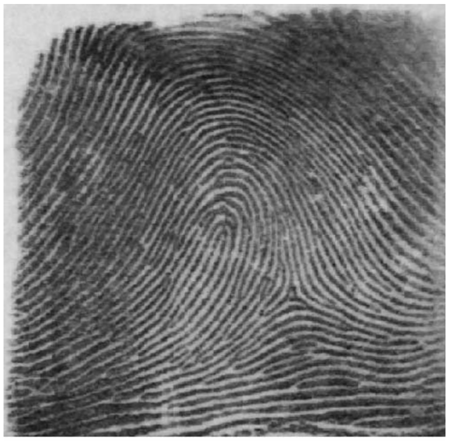


Fig. 15. Counterfeit image (Wong scheme). Vector quantization attack uses 8 × 8 blocks from 19 watermarked images.

Fig. 13 (top and center). In the third step, integrity and authenticity of the manipulated image is tested using the watermark detection algorithm. Output of the watermark detection step is seen in Figs. 12 and 13 (bottom). Numbers and shading indicate the lowest level signature verified, where darkest regions are not verified at any level. In Fig. 12, 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.

Effectiveness of the proposed algorithm against a VQ attack has been demonstrated in the second test case. As in [14], a database of fingerprint images has been used for this attack. The images of size 640 × 640 are first watermarked by Wong's original scheme and our hierarchical method, separately, using the minimum block sizes possible. The example illustrated here utilizes a private key implementation with a 64 bit MAC for which the minimum block sizes for Wong's scheme and the proposed scheme are 8 × 8 and 10 × 10, respectively. The slightly larger block size in the latter case arises because the LSB payload consists of not only the block signature but also the accumulated signatures from various levels of the hierarchy. While the original unwatermarked image is seen in Fig. 14, counterfeit images for Wong's original scheme and our hierarchical method are presented in Figs. 15 and 16. In both cases, 19 watermarked images are utilized for constructing vector quantization codebooks. Note that the VQ codebooks consist of 8 × 8 and 10 × 10 blocks corresponding to the minimum block sizes used by the watermarking algorithms. As a result of a successful attack, Fig. 15 is verified as authentic by Wong's scheme. Output of our hierarchical method indicates that the attack is indeed successful at the lowest level of the hierarchy. Signatures of all

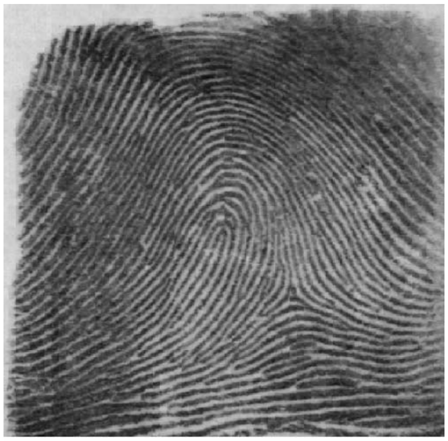


Fig. 16. Counterfeit image (Proposed hierarchical method). Vector quantization attack uses 10 × 10 blocks from 19 watermarked images.

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.

In this example, we assumed that the attacker uses the lowest level blocks, yet given sufficient resources larger level blocks may be used in the process. Even in that case, the attack will be detected on the next higher level, unless top level block, the image itself, isn't verified as a whole using the signature. In fact, the hierarchical method provides complete resilience against vector quantization counterfeiting attacks.

B. Localization Accuracy

In Wong's scheme, tamper detection is done on a block basis. Thus, tamper localization ability of the scheme is bounded by the block size used. On the other hand, as the signature of each block is inserted into the least significant bit-plane of the block, minimum block size is determined by the length of the signature used

$$\mathcal{L} \geq \mathcal{S} \tag{19}$$

where $\mathcal{S}$ is the length of the signature, and $\mathcal{L}$ is the tamper localization ability of the algorithm in pixels.

Similarly, localization ability of our method is bounded by the size of the blocks on the lowest level of the hierarchy. Nonetheless, the relation between the signature size $\mathcal{S}$ and localization ability $\mathcal{L}$ is not immediately obvious, since the LSB-plane of each block at the lowest level carries a larger payload then its signature. In a hierarchy of L levels, payload of such a block, and the localization ability $\mathcal{L}$ thereof, can be calculated as

$$\mathcal{L} = \sum_{l=0}^{L-1} 4^{-l}\mathcal{S} \leq \frac{4\mathcal{S}}{3}. \tag{20}$$

In particular, our algorithm slightly compromises localization ability in comparison with the original scheme, in order to gain increased robustness against VQ attacks. Despite the loss relative to Wong's scheme, fine granular localization (down to 10 × 10 blocks) can be achieved in practice.

C. Security Under Brute Force Attacks

Security of an algorithm may only be evaluated against known attacks. Two particular points of concern in fragile watermarking algorithms are forgery with brute-force attack and forgery with vector quantization attack. In the previous section we have demonstrated the effectiveness of our system against vector quantization counterfeiting.

In this section, we will compare the strength of our method with original scheme under brute-force attacks on digital signatures. As none of the methods specify a particular signature algorithm, we will assume that a signature of length n requires $f(n)$ trials for a successful forgery, and same length signatures are used for both of the methods. We further assume that minimum possible block sizes are used to partition an image of size A in both cases.

Number of signatures embedded in each case equals to the total number of blocks. In the original scheme the image is tiled using blocks of size n; thus the number of blocks and thereof signatures is:

$$N_{original} = \frac{A}{n}. \tag{21}$$

On the other hand, in the hierarchical scheme blocks of size $(4/3)n$ (ref. Section IV-B) tile the image. After higher level block signatures are included, total number of signatures will be:

$$N_{hierarchical} \simeq \sum_{l=0}^{L-1} 4^{-l}\frac{A}{\frac{4}{3}n} \simeq \frac{A}{n}. \tag{22}$$

That is, perfect forgery of an image using a brute-force attack will require roughly equal number of signature forgeries $[(A/n)f(n)$ trials] in each case.

D. Computational Complexity

A discretionary categorization of operations involved in the proposed approach will yield two classes; namely, digital signature computation(/verification) operations, and memory manipulations which prepare image data for these operations. Since the computational complexity of the latter group is negligible with respect to the first, from now on we will refer to the number of digital signature operations required as the *computational complexity* of the algorithm.

Given the above definition, let us analyze the complexity of the original scheme and the proposed approach. First, we will consider both schemes without the cropping detection functionality outlined in Section III-D.

The number of digital signatures present in each scheme is derived in Section IV-B, in the context of brute force attacks. As a result, it is established that both schemes require approximately the same number of digital signatures i.e., (A/n), and therefore (A/n) signature operations. The computational complexities of these schemes are thus approximately equal.

Now, let us consider the complexity of an hierarchical "sliding window" search explained in Section III-D. Without loss of generality, we assume that during embedding the image has been organized in an L level hierarchy with $O \times P$ blocks at the lowest level. On the lowest level, a "sliding window" search in a $O \times P$ neighborhood is sufficient to regain synchronization, i.e., at least one block is guaranteed to start in this neighborhood. Since block boundaries of different levels are aligned, once the synchronization is regained at some level, the search at any subsequent (higher) level need only consider possible groupings of the blocks obtained at the preceding (lower) level of the hierarchy, which is only a very small subset of all possible positionings. In particular, the "sliding window" of a higher level consists of a 2 × 2 neighborhood of lower level blocks. Therefore, in the worst case, regaining synchronization after an arbitrary cropping requires

$$C_{hierarchical} = O \times P + (L-1)2 \times 2 = OP + 4(L-1) \quad (23)$$

digital signature operations. Moreover, the average computational burden will be half of the complexity for the worst case scenario, if the cropping position is assumed to be uniformly distributed. Note that the corresponding search complexity in Wong's original scheme is rather small, because a single level search is sufficient, and furthermore for the same digital signature algorithm, Wong's scheme allows for smaller block-sizes than the scheme proposed here (Section IV-B).

In order to quantify values of the worst case search complexity encountered in practice, let us consider the fingerprint database example of Section IV-A, where original and proposed schemes operate on 8 × 8 and 10 × 10 blocks, respectively. In the latter case, the hierarchy is composed of seven levels. In this example, the proposed scheme requires 120 digital signature operations for cropping detection. Wong's original scheme, on the other hand, requires 64 signature operations, which is roughly half of what is required by the proposed scheme. Nonetheless, in either case the increase in overall complexity due to cropping detection is rather insignificant, 1% and 2%, respectively. Therefore, in typical images, the computational complexities of the two schemes with cropping detection are comparable.

V. Conclusion and Discussion

In this paper, we describe a new hierarchical fragile watermarking scheme based on the public key watermark by Wong [8]. The proposed method eliminates the vulnerabilities of the original scheme to VQ counterfeiting attack of Holliman and Memon [14]. 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.

The hierarchical scheme 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.

In this paper, the hierarchical watermarking scheme was described as applied to Wong's scheme. The method can, however, be readily applied to other block-wise independent fragile watermarking algorithms in order to thwart vector quantization counterfeiting attacks.

VQ based attacks on blockwise independent watermarking schemes hint at the existence of a trade-off between the accuracy of localization and the security/robustness (for semi-fragile schemes) of watermarking methods aimed at tamper localization. This seems analogous to the inherent trade-off between embedding distortion and capacity encountered with robust watermarking [17] and is worth investigating further.

References

[1] A. Menezes, P. van Oorchot, and S. Vanstone, *Handbook of Applied Cryptography*. Boca Raton, FL: CRC, 1997.

[2] R. L. Lagendijk, G. C. Langelaar, and I. Setyawan, "Watermarking digital image and video data," *IEEE Signal Processing Mag.*, vol. 17, pp. 20–46, Sept. 2000.

[3] F. Hartung and M. Kutter, "Multimedia watermarking techniques," *Proc. IEEE*, vol. 87, pp. 1079–1107, July 1999.

[4] M. D. Swanson, M. Kobayashi, and A. H. Tewfik, "Multimedia data-embedding and watermarking technologies," *Proc. IEEE*, vol. 86, pp. 1064–1087, June 1998.

[5] I. J. Cox and M. L. Miller, "A review of watermarking and the importance of perceptual modeling," *Proc. SPIE*, vol. 3016, Feb. 1999.

[6] G. L. Friedman, "The trustworthy digital camera: Restoring credibility to the photographic image," *IEEE Trans. Consumer Electron.*, vol. 39, pp. 905–910, Nov. 1993.

[7] M. Yeung and F. Mintzer, "An invisible watermarking technique for image verification," in *Proc. IEEE Int. Conf. Image Processing*, Santa Barbara, CA, Oct. 1997, pp. 680–683.

[8] P. W. Wong, "A public key watermark for image verification and authentication," in *Proc. IEEE Int. Conf. Image Processing*, Chicago, IL, October 4–7, 1998, pp. 425–429.

[9] J. Fridrich, M. Goljan, and A. C. Baldoza, "New fragile authentication watermark for images," in *Proc. IEEE Int. Conf. Image Processing*, Vancouver, BC, Canada, Sept. 10–13, 2000.

[10] C. W. Wu, D. Coppersmith, F. C. Mintzer, C. P. Tresser, and M. M. Yeung, "Fragile imperceptible digital watermark with privacy control," *Proc. SPIE, Security and Watermarking of Multimedia Contents I*, vol. 3657, Jan. 1999.

[11] D. Kundur and D. Hatzinakos, "Digital watermarking for telltale tamper proofing and authentication," *Proc. IEEE*, vol. 87, pp. 1167–1180, July 1999.

[12] J. Eggers and B. Girod, "Blind watermarking applied to image authentication," in *Proc. IEEE ICASSP*, Salt Lake City, UT, May 2001.

[13] S. Bhattacharjee and M. Kutter, "Compression tolerant image authentication," in *Proc. IEEE Int. Conf. Image Processing*, Chicago, IL, Oct. 1998.

[14] M. Holliman and N. Memon, "Counterfeiting attacks on oblivious block-wise independent invisible watermarking schemes," *IEEE Trans. Image Processing*, vol. 9, pp. 432–441, Mar. 2000.

[15] P. W. Wong and N. Memon, "Secret and public key authentication watermarking schemes that resist vector quantization attack," *Proc. SPIE*, vol. 3971, no. 40, Jan. 2000.

[16] J. Fridrich, "Security of fragile authentication watermarks with localization," *Proc. SPIE*, vol. 4675, no. 75, Jan. 2002.

[17] P. Moulin and J. A. O'Sullivan, "Information-theoretic analysis of information hiding," Pre-print, http://www.ifp.uiuc.edu/moulin/Papers/ITthiding99.ps.gz, Sept. 1999.

Mehmet Utku Celik (S'98) received the B.Sc. degree in electrical and electronic engineering in 1999 from Bilkent University, Ankara, Turkey and the M.Sc. degree in electrical and computer engineering in 2001 from the University of Rochester, Rochester, NY, where he is currently pursuing the Ph.D. degree.

Currently, he is a Research Assistant in the Electrical and Computer Engineering Department, University of Rochester. His research interests include digital watermarking and data hiding—with emphasis on multimedia authentication—image and video processing, and cryptography.

Mr. Celik is a member of the ACM and the IEEE Signal Processing Society.

Gaurav Sharma (S'88–M'96–SM'00) received the B.E. degree in electronics and communication engineering from University of Roorkee, India in 1990, the M.E. degree in electrical communication engineering from the Indian Institute of Science, Bangalore, in 1992, and the M.S. degree in applied mathematics and Ph.D. degree in electrical and computer engineering from North Carolina State University (NCSU), Raleigh, in 1995 and 1996, respectively.

From 1992 to 1996, he was a Research Assistant at the Center for Advanced Computing and Communications in the Electrical and Computer Engineering Department at NCSU. Since 1996, he has been a Member of Research and Technical Staff at Xerox Corporation's Digital Imaging Technology Center, Webster, NY. He is also involved in teaching in an adjunct capacity at the Electrical Engineering Department at the Rochester Institute of Technology, Rochester, NY. His research interests include image security and watermarking, color science and imaging, signal restoration, and halftoning.

Dr. Sharma is a member of Sigma Xi, Phi Kappa Phi, Pi Mu Epsilon, and is the vice president for the Rochester Chapter of the IEEE Signal Processing Society.

Eli Saber (S'91–M'96–SM'00) received the B.S. degree in electrical and computer engineering from the University of Buffalo, Buffalo, NY, in 1988, and the M.S. and Ph.D. degrees in electrical and computer engineering from the University of Rochester, Rochester, NY, in 1992 and 1996 respectively.

He joined Xerox in 1988 and is currently a Product Development Scientist and Manager heading the Image Science, Analysis, and Evaluation area in the Print Engine Development Unit. He is an Adjunct Faculty Member at the Electrical and Computer Engineering Departments of the University of Rochester and the Rochester Institute of Technology responsible for teaching graduate coursework in signal, image and video processing and performing research in digital libraries and watermarking. His research interests include color image processing, image/video segmentation and annotation, content-based image/video analysis and retrieval, computer vision, and watermarking. He holds a number of conference and journal publications in the field of signal, image, and video processing.

Dr. Saber was the recipient of the Gibran Khalil Gibran Scholarship and of several prizes and awards for outstanding academic achievements from 1984 to 1988, as well as the Quality Recognition Award in 1990 from The Document Company, Xerox. He is a member of the Electrical Engineering Honor Society, Eta Kappa Nu, and the Imaging Science and Technology Society.

Ahmet Murat Tekalp (S'80–M'84–SM'91) received the M.S. and Ph.D. degrees in electrical, computer, and systems engineering from Rensselaer Polytechnic Institute (RPI), Troy, New York, in 1982 and 1984, respectively.

From December 1984 to August 1987, he was a Research Scientist at Eastman Kodak Company, Rochester, New York. He joined the Electrical and Computer Engineering Department, University of Rochester, Rochester, NY, in September 1987, where he is currently an endowed Distinguished Professor. His current research interests are in the area of digital image and video processing, including image restoration, video segmentation, object tracking, content-based video description, and protection of digital content. At present, he is the Editor-in-Chief of the *EURASIP Journal on Image Communication*. He was associate editor for the *Journal of Multidimensional Systems and Signal Processing* (1994–1999). He was an area editor for the Academic Press Journal Graphical Models and Image Processing (1995–1998). He was also on the editorial board of the Academic Press Journal Visual Communication and Image Representation (1995–1999). He authored *Digital Video Processing* (Englewood Cliffs, NJ: Prentice-Hall, 1995). He holds five U.S. patents. His group contributed technology to the ISO/IEC MPEG-4 and MPEG-7 standards.

Dr. Tekalp received the NSF Research Initiation Award in 1988, was named as Distinguished Lecturer by IEEE Signal Processing Society in 1998, and was awarded a Fulbright Senior Scholarship in 1999. He has chaired the IEEE Signal Processing Society Technical Committee on Image and Multidimensional Signal Processing (Jan. 1996–Dec. 1997). He has served as an Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING (1990–1992), IEEE TRANSACTIONS ON IMAGE PROCESSING (1994–1996). He was appointed as the Technical Program Chair for the 1991 IEEE Signal Processing Society Workshop on Image and Multidimensional Signal Processing, the Special Sessions Chair for the 1995 IEEE International Conference on Image Processing, and the Technical Program Co-Chair for IEEE ICASSP 2000 in Istanbul, Turkey. He is the General Chair of IEEE International Conference on Image Processing (ICIP) 2002 at Rochester, NY. He is the Founder and First Chairman of the Rochester Chapter of the IEEE Signal Processing Society. He was elected as the Chair of the Rochester Section of IEEE in 1994–1995.