Digital Watermarking for Telltale Tamper Proofing and Authentication

DEEPA KUNDUR, STUDENT MEMBER, IEEE, AND DIMITRIOS HATZINAKOS, SENIOR MEMBER, IEEE

Invited Paper

In this paper, we consider the problem of digital watermarking to ensure the credibility of multimedia. We specifically address the problem of fragile digital watermarking for the tamper proofing of still images. Applications of our problem include authentication for courtroom evidence, insurance claims, and journalistic photography.

We present a novel fragile watermarking approach which embeds a watermark in the discrete wavelet domain of the image by quantizing the corresponding coefficients. Tamper detection is possible in localized spatial and frequency regions. Unlike previously proposed techniques, this novel approach provides information on specific frequencies of the image that have been modified. This allows the user to make application-dependent decisions concerning whether an image, which is JPEG compressed for instance, still has credibility. Analysis is provided to evaluate the performance of the technique to varying system parameters. In addition, we compare the performance of the proposed method to existing fragile watermarking techniques to demonstrate the success and potential of the method for practical multimedia tamper proofing and authentication.

Keywords—Authentication, data hiding, digital watermarking, steganography, telltale tamper proofing.

I. Introduction

Research in the area of digital watermarking has focused primarily on the design of robust techniques for the copyright protection of multimedia data. In such methods a watermark is imperceptibly embedded in a host signal such that its removal using common distortions on the marked signal is difficult without degrading the perceptible data content itself. Watermarking can also be used to address the equally important, but underdeveloped, problem of tamper proofing.

As a great deal of multimedia is stored in digital format, it has become easier to modify or forge information using widely available editing software. In fact, almost all

Manuscript received February 27, 1998; revised December 1, 1998. This work was supported in part by the Natural Sciences and Engineering Research Council (NSERC) of Canada and by the Communications and Information Technology Ontario (CITO).

The authors are with the Department of Electrical and Computer Engineering, University of Toronto, Toronto, Ont. M5S 3G4 Canada (e-mail: deepa@comm.toronto.edu; dimitris@comm.toronto.edu).

Publisher Item Identifier S 0018-9219(99)04950-6.

published imagery is edited to some extent using computerbased tools. A problem arises when the possibly tampered data are to be used as evidence; in such situations, the multimedia data must be credible. By "credible" we mean that the signal source is authentic and that the information content in the signal has not been modified in transit to its destination.

In this paper, we present a technique for signal tamper proofing. Previously proposed methods for images [1]–[5] place the watermark in the spatial domain of the signal; they provide information on the spatial location of the changes but fail to give a more general characterization of the type of distortion applied to the signal. In contrast, our scheme places the watermark in the discrete wavelet domain, which allows the detection of changes in the image in localized spatial and frequency domain regions. This gives our approach the versatility to detect and help characterize signal modifications from a number of distortions such as substitution of data, filtering, and lossy compression. In addition, we embed the mark by quantizing the coefficients to a prespecified degree, which provides the flexibility to make the tamper-proofing technique as sensitive to changes in the signal as desired. We call such a method a telltale tamper-proofing scheme.

The main objectives of this paper are:

- 1) to introduce a set of well-defined goals for a telltale tamper-proofing scheme;
- to present a novel tamper-proofing and authentication technique which provides more complete information on how the image is modified;
- to demonstrate the potential of tamper-proofing methods through implementations of our method and existing techniques;
- to provide a comparative study of the strengths and limitations of the proposed and existing tamperproofing methods.

In Section II we define the specific problems we address in this paper and provide a review of existing techniques for the tamper proofing of images. We propose and intro-

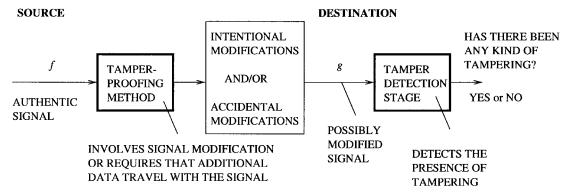


Fig. 1. The traditional tamper-proofing problem.

duce a set of objectives for the novel problem of telltale tamper proofing. The proposed technique is developed and analyzed using concepts from signal detection theory in Section III. Implementation issues are discussed in Section IV. Simulation results and comparisons of the performance of the technique to previously proposed methods are provided in Section V, followed by concluding statements in Section VI.

II. PROBLEM FORMULATION

A. Tamper Proofing Versus Telltale Tamper Proofing

The problem we address is that of the telltale tamper-proofing of multimedia signals for authentication. The traditional problem of tamper-proofing can be stated as follows. Consider the existence of an original or authentic digital multimedia signal f. Given a signal g, which is a possibly modified version of f, determine to a high degree of probability, whether g=f without explicit knowledge of the original signal f.

Thus, if it can be shown that g is equal to f almost for certain, then the signal g is considered to be *credible*. There are two basic stages to the process of tamper proofing. In the first stage (at the source) the original signal is passed through a hash function to produce a piece of data separate from the signal¹; these data are then used in the second stage (at the receiver) to verify that the received image has not been modified. Alternatively, it has been shown that the verification data can be directly embedded imperceptibly into the signal [2]. These data are extracted from the signal itself in the second stage to check for tampering. Fig. 1 gives an overview of the tamper-proofing problem.

Several approaches have been recently proposed to address the issue of tamper proofing. In [1], Friedman describes a "trustworthy digital camera" in which a digital camera image is passed through a hash function and then is encrypted using the photographer's private key to produce a piece of authentication data separate from the image. These data are used in conjunction with the image to ensure that no tampering has occurred. Specifically, the photographer's public key is used to decrypt the hashed original image and the result is compared to the hashed version of the received

image to ensure authentication. In [2], Walton proposes a technique in which a separate piece of data is not required for authentication. The method requires the calculation of the checksums of the seven most significant bits of the image (or a transformed version of the image), so that they may be embedded into randomly selected least significant bits. The major disadvantage of the techniques in [1] and [2] is that they produce a dichotomous result (i.e., "yesor-no" solution) to the question of tampering; it is not straightforward to determine how the image is tampered which makes the scheme highly susceptible to random bit errors during data transmission.

For the tamper proofing of multimedia signals there is an additional issue of incidental distortions the signal may undergo due to compression, enhancement, or transmission errors. For many applications, such transformations of the signal are necessary and still maintain the integrity of the signal information. Thus, in this paper, we consider the more practical issue of identifying whether or not the tampering on the signal, if any, has an effect on its "credibility."

A few techniques which attempt to address this problem have been proposed in the literature. In [3], Schneider and Chang propose a method for content-based image verification in which they define a continuous interpretation of the concept of authenticity which measures the closeness of specific features of a possibly modified image to the original one. The procedure is comprised of three stages in which: 1) the relevant signal content is extracted; 2) the results of stage 1) are hashed to reduce size; and 3) the result of stage 2) is encrypted with the author's private key. The image content extraction could be localized histogram information, discrete cosine transform (DCT) coefficients, or edge information. The advantage of the method is that signals that undergo incidental distortions can still be deemed credible. However, the process of selecting the image content extraction functions used in stage 1) is not straightforward for a given application.

Wolfgang and Delp in [4] proposed a fragile watermarking technique involving the addition of two-dimensional m sequences. They define a nonbinary test statistic based on the inner product of the m sequence and the image which gives a relative measure of the tampering of a particular image block. The major disadvantage is that it

¹These data can also be an encrypted author ID independent of the signal.

is possible to modify the data without disturbing the lower significant bits which contains the verification information. Similarly in [5], Yeung and Mintzer discuss a digital watermarking technique which tries to detect the modification of individual pixels. The technique requires the use of a look-up table (LUT) which maps image colors to binary numbers. The original image pixel colors are modified such that the associated binary numbers determined from the LUT equal the watermark bit values. Although the techniques in [4] and [5] give information about spatially localized changes in the image, they do not provide more explicit information on how the image is tampered. For example, if the image is innocently lossy compressed for convenience, then the entire image may appear tampered and its usefulness ignored.

We argue that traditional authentication approaches for data are not well suited for images, sound, and video; to be practically useful a tamper-proofing technique must not only detect the presence of modifications in a signal but should also provide information helpful to characterize the distortions. A telltale tamper proofing method must be able to do the following:

- 1) indicate with high probability that some form of tampering has or has not occurred;
- 2) provide a measure of the relative degree of distortion of the signal;
- 3) characterize the type of distortion, such as filtering, compression or replacement, without access to the original host signal or any other signal-dependent information; it should be possible to detect changes due to compression or random bit errors and make application-dependent decisions concerning whether or not the signal still has credibility;
- 4) validate the signal and authenticate the source without requiring the maintenance and synchronization of additional data separate from the signal.

There has been a recent trend toward addressing the problems of tamper proofing and authentication using a digital watermarking approach. The attraction of such an approach is that no additional data are required for signal verification. In addition, the verification information is discretely watermarked which adds an additional level of security against attacks to modify both the signal and the verification data. In the next section, we discuss the digital watermarking problem.

B. The Digital Watermarking Approach

Traditionally, digital watermarking has been used to embed author and copyright identification into a multimedia signal [6]–[11]. The watermark must be retained in the signal even under intentional signal distortion attacks to remove it. In contrast, fragile watermarking refers to the process of marking a signal such that any modification causes the extracted mark to be different than the original which indicates that tampering has taken place.

We briefly discuss some terminology and requirements for a successful fragile watermarking method. We assume without loss of generality that the signal to be marked is a still image. A fragile watermark w is defined as a signal (which is often a randomly generated binary stream), containing information used to assess whether an image was modified. The watermark is considered to be fragile because it is embedded in a way such that any slight modification of the resulting image will distort the watermark as well. The embedding procedure involves modifying a host image to reflect the information content in w. The modification must be imperceptible in the sense that the owner and recipient of the signal show no preference to the information content in either the original or marked signal. Watermark extraction is the process of detecting the presence of watermark information in a given image and is performed to recover the mark and to assess whether tampering has been performed.

Some recent work in fragile watermarking [2], [4], [5] has demonstrated the potential of the approach. We specifically define the problem of fragile watermarking for the application of telltale tamper-proofing as follows. Given a digital multimedia signal f and a digital watermark w, embed w into f by imperceptibly modifying f to produce a tamper-proofed signal z such that:

- 1) the watermark w can be extracted from z without requiring explicit knowledge of f;
- 2) if the information content in z is unmodified, then the extracted watermark \tilde{w} exactly matches w;
- 3) if z is modified, then \tilde{w} is different from the embedded with a probability vanishing close to one;
- the differences between the embedded and extracted watermarks provide useful information to assess whether the signal modification maintains or destroys credibility.

We present a watermarking technique which attempts to address the above criteria.

III. PROPOSED TECHNIQUE

A. General Approach

Our technique is described in the context of watermarking still images, but it also works for general multimedia signals. We make use of the discrete wavelet domain opposed to spatial or DCT domains to embed the watermark because it provides both a simultaneous spatial localization and a frequency spread of the watermark within the host image. The localization of the watermark gives the ability to identify distinct regions of the watermarked image which have undergone tampering and the global spreading of the mark makes it sensitive to large-scale signal distortions. We argue that characterizing the modifications in terms of localized space-frequency distortions is more effective and practical for tamper proofing than attempting to parameterize the distortion. Parametric models can be highly inaccurate in estimating a wide class of image transformations and are often costly to compute for larger images.

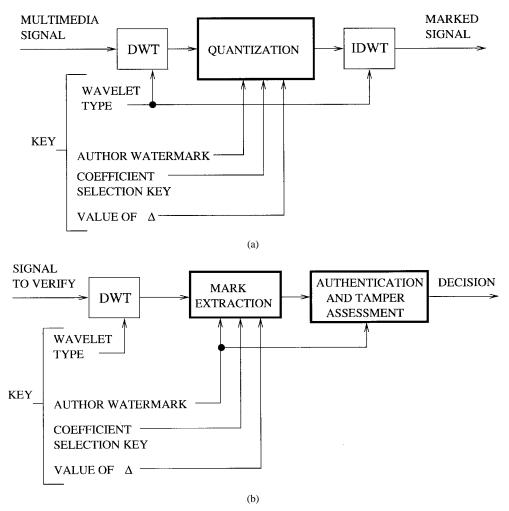


Fig. 2. Proposed telltale tamper-proofing approach: (a) embedding process and (b) tamper assessment process.

The fundamental advantage of our technique lies in its ability to detect, with high probability, the spatial and frequency components of the image which are untampered and, hence, still credible. We embed the mark by quantizing the coefficients to a prespecified degree which provides the flexibility to make the tamper-proofing technique as sensitive to changes in the signal as desired. The general scenario is shown in Fig. 2(a). A validation key comprised of the author's watermark, a coefficient selection key (which we describe later), the quantization parameter Δ , and possibly the specific mother wavelet function are necessary for embedding and extracting the mark. The watermark can be an encrypted version of the author identification which is used to establish sender authenticity. There are three main stages to the watermark embedding procedure.

In the first stage, we compute the Lth-level discrete wavelet decomposition of the host image to produce a sequence of three L detail images, corresponding to the horizontal, vertical, and diagonal details at each of the L resolution levels, and a gross approximation of the image at the coarsest resolution level. The value of L is user defined. We denote the kth detail image component at the lth resolution level of the host by $f_{k,l}(m,n)$, where k=h,v,d (which stands for "horizontal," "ver-

tical," and "diagonal" detail coefficients, respectively), l = $1, \ldots, L$ and (m, n) is the particular spatial location index at resolution l. The gross approximation is represented by $f_{a,L}(m,n)$ where the subscript a is used instead of k to denote "approximation." In the second stage, we embed the watermark bit stream by modifying selected wavelet coefficients. Specifically, to embed a binary watermark of length N_w denoted w(i), $i = 1, 2, ..., N_w$, a user-defined coefficient selection key ckey(i), $i = 1, 2, ..., N_w$ is employed. The particular wavelet coefficient at which to embed the *i*th watermark bit w(i) is given by ckey(i). Each element of ckey is distinct so that two bits are not marked at the same location, causing an ambiguity or error. In addition, the selection of the coefficients is random and well spread spatially and throughout each resolution level to be able to assess changes to these image components. In the simulations for this paper, ckey(i) was generated by randomly selecting a coefficient from the set $\{f_{h,l}(m,n), f_{v,l}(m,n), f_{d,l}(m,n)\}$ for each l and (m, n). Thus, one detail coefficient at each resolution and spatial location was marked. The binary watermark was also randomly generated using a uniform distribution and was set to be the same length as ckey. The watermark bit w(i) is embedded into the coefficient ckey(i) through

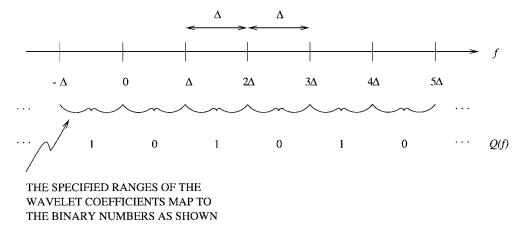


Fig. 3. The quantization function. Each possible real value of the detail coefficient has an associated binary number given by $Q(\cdot)$.

an appropriate quantization procedure. The specifics of the quantization are discussed in the next section. In the final stage, the corresponding Lth-level inverse wavelet transform of the marked image components is computed to form the tamper-proofed image.

Watermark extraction on a given image is performed as shown in Fig. 2(b). The Lth-level discrete wavelet transform (DWT) is applied to the given image and the coefficient selection key ckey(i) is used to determine the marked coefficients. A quantization function $Q(\cdot)$ (also discussed in the next section) is applied to each of these coefficients to extract the watermark values. For authentication, the author's public key is applied to the extracted watermark to obtain the author identification code. Almost any tampering of the image will cause the authentication procedure to fail as the decryption procedure is highly sensitive to changes in the watermark. Thus, authentication is possible only if the extracted watermark is identical to the embedded. If public key authentication fails, then we employ tamper assessment to determine the credibility of the modified multimedia content.

To assess the extent of tampering, we compute the following function which we call the tamper assessment function (TAF)

$$TAF(w, \tilde{w}) = \frac{1}{N_w} \sum_{i=1}^{N_w} w(i) \oplus \tilde{w}(i)$$
 (1)

where w is the true author watermark, \tilde{w} is the extracted mark, N_w is the length of the watermark, and \oplus is the exclusive-OR (XOR) operator. The value of $TAF(w, \tilde{w})$ ranges between zero and one. To determine image modifications for specific frequencies and/or spatial regions the watermark can be extracted from the corresponding marked wavelet coefficients alone.

The presence of tampering is determined if $TAF(w, \tilde{w}) \geq \mathcal{T}$, where $0 \leq \mathcal{T} \leq 1$ is prespecified threshold. If $TAF(w, \tilde{w}) < \mathcal{T}$, then the modifications on the image are considered to be incidental and negligible. For higher security applications, \mathcal{T} can be set to be smaller. The magnitude of $TAF(w, \tilde{w})$ can be used to assess the extent of tampering. We show in Section V that if JPEG

compression is applied to an image, the method can assess that most of the changes have occurred to the details at the higher resolution levels. If a part of the image has been replaced/changed in addition to compression, the watermark in the lower resolutions will not remain the same. Hence, the lower resolution image can be authenticated. In addition, when filtering is applied to an image the technique can assess the frequency regions most tampered with [12].

B. Details of the Quantization Process

For an arbitrary wavelet transform, the detail coefficients $\{f_{k,l}(m,n)\}$ are real numbers. We perform quantization on the wavelet coefficients in the following manner. Every real number is assigned a binary number, as shown in Fig. 3. We denote this function by $Q(\cdot)$ which maps the real number set to $\{0, 1\}$. Specifically

$$Q(f) = \begin{cases} 0, & \text{if } r\Delta \le f < (r+1)\Delta \text{ for } r = 0, \pm 2, \pm 4, \dots \\ 1, & \text{if } r\Delta \le f < (r+1)\Delta \text{ for } r = \pm 1, \pm 3, \pm 5, \dots \end{cases}$$
(2)

where Δ is a positive real number called the quantization parameter and is shown in Fig. 3. The following assignment rule is used to embed the watermark bit w(i) into the selected coefficient ckey(i). We denote the coefficient selected by ckey(i) as $f_{k,l}(m,n)$.

- 1) If $Q(f_{k,l}(m,n)) = w(i)$, then no change in the coefficient is necessary.
- 2) Otherwise, change $f_{k,l}(m,n)$ so that we force $Q(f_{k,l}(m,n)) = w(i)$, using the following assignment:

$$f_{k,l}(m, n) = \begin{cases} f_{k,l}(m, n) + \Delta & \text{if } f_{k,l}(m, n) \le 0\\ f_{k,l}(m, n) - \Delta & \text{if } f_{k,l}(m, n) > 0 \end{cases}$$
(3)

where Δ is the same parameter as in Fig. 3 and (2), and := is the assignment operator.

The nature of the assignment in (3) has been experimentally found to change the image with the least visual degradation for a given magnitude of Δ . The parameter Δ is user

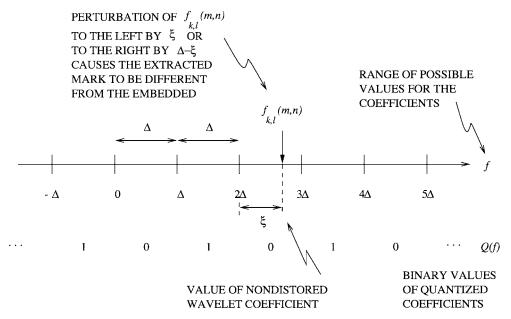


Fig. 4. Effect of noise on the extracted watermark bit. Perturbation of the wavelet coefficient from tampering can cause the extracted watermark to be different than the embedded.

defined and is set to establish an appropriate sensitivity to changes in the image. A smaller value of Δ will make the quantization process of the second stage finer and hence makes minor changes in the image easier to detect.

It is assumed that the specific wavelet transform used is unknown to make forgery difficult. If the wavelet transform were known, it would be possible for a tamperer to apply it to any arbitrary image and quantize the coefficients using the knowledge of $Q(\cdot)$ in the same way in which it appears in the original watermarked image so that the forgery appears authentic. We discuss how to overcome this handicap in Section IV with the use of an image-dependent quantization key.

C. Performance Analysis

In this section, we assess the performance of our general technique as a function of the system parameters. We concentrate on two types of degradations on a given region of the image.

- 1) Mild distortion, in which we model the degradation on the associated wavelet coefficients as additive noise with a probability density function (pdf) with rapidly decaying tails. We specifically model the distortion on the wavelet coefficients as zero mean additive Gaussian noise (AGN) with variance σ_v^2 . We assume that σ_v/Δ is "small," such that the Gaussian pdf rapidly decays. Examples of image distortions which can fall under this category are mild filtering and JPEG compression.
- 2) Severe distortion, in which the degradation is assumed to be additive noise consisting of heavy tails (i.e., σ_v/Δ is "large") and the value of the distorted wavelet coefficients becomes difficult to predict given

the true values. In fact, we consider that the probability of false watermark detection in such a degraded coefficient be 1/2. Heavy linear or nonlinear filtering, random bit errors, and image region substitution fall under this class of distortions.

We consider each type of distortion in turn to assess the performance of our method. We evaluate the effectiveness of the approach to tamper proofing by introducing a measure we call the tamper sensitivity function (TSF). We define this as the probability that tampering is detected for $\mathcal{T}=0$ given that N coefficients in the wavelet domain are modified.

1) Sensitivity of the Technique to Mild Distortion:

To assess the TSF for mild distortion we model the effects of the image degradation on a given wavelet coefficient as

$$\hat{f}_{k,l}(m,n) = f_{k,l}(m,n) + v_{k,l}(m,n) \tag{4}$$

where $f_{k,l}(m,n)$ is the undistorted wavelet coefficient, $\hat{f}_{k,l}(m,n)$ is the distorted coefficient, and $v_{k,l}(m,n)$ is the associated zero mean AGN with variance σ_v^2 . Without loss of generality we assume that $Q(f_{k,l}(m,n))=0$. Fig. 4 shows how the additive noise can perturb the wavelet coefficient such that $Q(f_{k,l}(m,n))=1$ so that the extracted watermark is different from that embedded.

The probability of false negative p_{fn} for the tampering of a given coefficient (i.e., the probability that tampering is not detected in a particular wavelet coefficient) is given by the probability that $Q(\hat{f}_{k,l}(m,n)) = 0$. That is

$$p_{fn} \stackrel{\Delta}{=} P\Big\{Q(\hat{f}_{k,l}(m,n)) = 0\Big\}. \tag{5}$$

Given that σ_v/Δ is small, we can neglect the probability of $|v_{k,l}(m,n)| > \Delta$ and we make the following approximation:

$$p_{fn} \approx P\{-\xi < v_{k,l}(m,n) < \Delta - \xi\} \tag{6}$$

²The tails of a pdf refer to the behavior of the function as the independent variable approaches infinity and negative infinity.

where ξ is the relative distance of the wavelet coefficient from one of the range boundaries of $Q(\cdot)$ as shown in Fig. 4. We simplify the expression for p_{fn} as follows by using the assumption that $v_{k,l}(m,n)$ is zero mean AGN

$$p_{fn} \approx P\{-\xi < v_{k,l}(m,n) < 0\}$$

$$+ P\{0 < v_{k,l}(m,n) < \Delta - \xi\}$$

$$= P\{0 < v_{k,l}(m,n) < \xi\}$$

$$+ P\{0 < v_{k,l}(m,n) < \Delta - \xi\}$$

$$= \operatorname{erf}\left(\frac{\xi}{2\sigma_{v}}\right) + \operatorname{erf}\left(\frac{\Delta - \xi}{2\sigma_{v}}\right)$$
(9)

where $erf(\cdot)$ is the traditional error function given by [13]

$$\operatorname{erf}(\xi) = \frac{2}{\sqrt{\pi}} \int_0^{\xi} e^{-t^2} dt.$$
 (10)

On average, ξ is evenly distributed between zero and Δ for an arbitrary image and wavelet transform. Therefore, the expected probability of a false negative of a degraded coefficient is given by

$$\overline{p}_{fn} = \frac{1}{\Delta} \int_0^{\Delta} \left[\operatorname{erf} \left(\frac{\xi}{2\sigma_v} \right) + \operatorname{erf} \left(\frac{\Delta - \xi}{2\sigma_v} \right) \right] d\xi \quad (11)$$

$$= \frac{2}{\Delta} \int_0^{\Delta} \operatorname{erf} \left(\frac{\xi}{2\sigma_v} \right) d\xi. \quad (12)$$

The TSF for mild distortion is given by the probability that at least one extracted watermark bit differs from the corresponding embedded bit

$$TSF_{mild} = 1 - P \begin{cases} all \ N \text{ modified coefficients produce} \\ false \text{ negative tampering results} \end{cases}$$
(13)

$$=1-\overline{p}_{fn}^{N}\tag{14}$$

$$=1 - \left[\frac{2}{\Delta} \int_0^\Delta \operatorname{erf}\left(\frac{\xi}{2\sigma_v}\right) d\xi\right]^N \tag{15}$$

where $\sigma_v/\Delta \ll 1$. Equation (15) gives us the average probability of tamper detection for $\mathcal{T}=0$ given that N wavelet coefficients have been modified. We see that the value of Δ dictates the sensitivity of the technique to image tampering. The value of the TSF increases monotonically with decreasing Δ ; hence, the smaller the value of Δ , the more sensitive the tamper detection which confirms our intuition. The value of Δ is user-defined so that the technique is flexible for a variety of applications. Equation (15) also reveals that there is a geometric increase in the change in the TSF as N is increased.

We next determine the probability of tamper detection for arbitrary $\mathcal{T}>0$ given that N coefficients are modified

$$P_{td}^{N} = P\{TAF(w, \, \tilde{w}) \ge T\}$$
 (16)

$$=P\left\{\sum_{i=1}^{N_w}w(i)\oplus\tilde{w}(i)\geq\mathcal{T}N_w\right\}. \tag{17}$$

The right-hand side of (17) is equivalently the probability that there are at least $\lceil \mathcal{T} N_w \rceil$ differences in the embedded

and extracted watermark bits. Using this interpretation, we conclude that [14]

$$P_{td}^{N} = \sum_{k=\lceil TN_{w} \rceil}^{N} \binom{N}{k} (1 - \overline{p}_{fn})^{k} \overline{p}_{fn}^{N-k}$$

$$= \sum_{k=\lceil TN_{w} \rceil}^{N} \binom{N}{k} \left[1 - \frac{2}{\Delta} \int_{0}^{\Delta} \operatorname{erf} \left(\frac{\xi}{2\sigma_{v}} \right) d\xi \right]^{k}$$

$$\cdot \left[\frac{2}{\Delta} \int_{0}^{\Delta} \operatorname{erf} \left(\frac{\xi}{2\sigma_{v}} \right) d\xi \right]^{N-k}$$
(19)

where N_w is the length of the extracted watermark, $\lceil \cdot \rceil$ is the ceiling operator, and it is assumed that $\sigma_v/\Delta \ll 1$. Equation (19) provides a relationship between the value of the threshold \mathcal{T} and the probability of tamper detection given that N wavelet coefficients have been mildly distorted. This probability can be set arbitrarily high by reducing both Δ and \mathcal{T} for a given σ_v .

2) Sensitivity of the Technique to Severe Distortion: As discussed in the beginning of Section III-C, we assume that the extracted mark is essentially independent of the embedded watermark value for severe distortion so that $\overline{p}_{fn}=1/2$. Performing a similar analysis to the mild distortion case, the TSF is given by

$$TSF_{severe} = 1 - \overline{p}_{fn}^{N}$$
 (20)

$$=1-\left(\frac{1}{2}\right)^N. \tag{21}$$

Since the distortion is unpredictable, (21) is independent of Δ , but there still remains a geometric relationship with N. The average probability of tamper detection for T>0 given that N wavelet coefficients are severely distorted is computed to be

$$P_{td}^{N} = \frac{1}{2^{N}} \sum_{k=\lceil T N_{tot} \rceil}^{N} \binom{N}{k} \tag{22}$$

where N_w is the length of the watermark extracted from a given region. The value of the threshold \mathcal{T} can be decreased to increase the probability of tamper detection.

When a geometric transformation such as shearing or rescaling is applied to a marked image, the locations of the fragile watermark bits become "unsynchronized." Therefore, we can model this distortion as severe since the probability of a false negative is essentially 1/2 as a watermark bit is extracted from a completely different location in the image and its value is unpredictable. Our scheme, as proposed in this paper, cannot be used to estimate the particular geometric transformation applied to the image; the tampering is merely detected.

IV. IMPLEMENTATION ISSUES FOR TAMPER PROOFING

A. The Algorithm

In this section, we discuss the major implementation issues of realizing our telltale tamper-proofing technique and our strategies to overcome them. We present the specific algorithm implemented. The two main obstacles in implementing the method are its numerical sensitivity and its susceptibility to forgery.

1) Numerical Sensitivity: The fragile watermarking of images is somewhat different than robust watermarking because the design of the technique must be intrinsically sensitive to detect tampering. Existing fragile watermarking methods deal with the addition of integers to the spatial domain pixels of the image [2], [4], [5]. Our proposed method involves embedding the watermark in the wavelet domain. When the marked wavelet coefficients are modified and the inverse DWT is applied, the resulting marked image pixels must be rounded to integer values to form a digital image. This rounding operation is an image modification that may cause the watermark in the marked image to differ from the original due to numerical sensitivity.

To avoid these numerical difficulties, we propose an algorithm in which the changes to the wavelet coefficients guarantee integer changes in the spatial domain. We make use of the Haar wavelet transform, in which the coefficients at each resolution level l are rational numbers of the form $r/2^l$ where $r \in \mathbb{Z}$. We modify the coefficients by adding or subtracting a multiple of 2^l . This specific type of quantization guarantees that inverse DWT produces an image with integer pixel values; no rounding, which may jeopardize the accuracy of the method, is necessary.

We use the following modified quantization function to embed the watermark such that $Q_{\delta}(f_{k,l}(m,n))$ is equal to the watermark bit value:

$$Q_{\delta}(f) = \begin{cases} 0, & \text{if } \left\lfloor \frac{f}{\delta 2^{l}} \right\rfloor \text{ is even} \\ 1, & \text{if } \left\lfloor \frac{f}{\delta 2^{l}} \right\rfloor \text{ is odd} \end{cases}$$
 (23)

where δ is a prespecified positive integer and $\lfloor \cdot \rfloor$ is the floor function.

2) Susceptibility to Forgery: As discussed in Section III-B, knowledge of the specific wavelet transform used to embed the watermark can jeopardize the security of the method. However, during implementation, we make exclusive use of the Haar wavelet, which is a disclosed detail of the algorithm. To combat this, we introduce an image-dependent key called the quantization key $qkey(i) \in \{0,1\}$. The value of this key for each index i is a function of a localized component of the image.

The purpose of the quantization key is to make the forgery of an untampered image virtually impossible without knowledge of qkey(i). Instead of embedding the watermark w(i) directly into the wavelet coefficients, we embed $w(i) \oplus qkey(i)$ where qkey(i) is dependent on the image. If we wanted to make the tamper proofing especially sensitive to changes in horizontal edges of the image, then the value of qkey could be a function of $f_{h,l}(m,n)$. Similarly, if we wanted the technique to indicate changes in the mean value of the image, then qkey can be dependent on localized averages of the image intensity. The introduction of qkey improves security against forgery and provides the flexibility to monitor specific changes to the image.

Algorithmic forms of the watermark embedding, and extraction and tamper assessment routines are provided in Tables 1 and 2, respectively. The choice of the user-defined parameters are discussed in Section V.

B. Key Features of the Algorithm

We discuss and review the main characteristics of the technique which distinguish it from previously proposed methods for watermarking.

- Our technique differs from existing fragile watermarking techniques in that the mark is embedded in the discrete wavelet domain. This allows information concerning the frequencies of the image that have undergone tampering and their relative degree of distortion.
- 2) There is a relationship between the value of the maximum wavelet decomposition level L and the visibility of the watermark. Given that $1 \le K \le 3$ detail coefficients are marked per spatial location at a particular resolution, there can be a change in any image pixel of at most KL. The larger the value of L the more localized the information that is extracted concerning changes to lower frequencies of the image. Thus, there exists a tradeoff between the visibility of the mark and the ability to detect changes in lower image frequencies. Analogously, increasing the value of K can provide additional information about tampering such as the possibility of directional filtering, but this increases the chance of visibility.
- 3) The quantization key provides the flexibility to make the technique more or less sensitive to certain distortions. For example, if we wish to detect changes in the mean value of each 8 × 8 block of the image, then *qkey* can depend directly on this quantity so that any change in the mean will scramble *qkey* and hence cause the extracted watermark value to differ from the embedded with high probability. It should be noted that the presence of *qkey* maintains the integrity of the tamper-proofing scheme against forgery even under the condition that the coefficient sensitivity function *ckey* is disclosed.

These properties make the method appealing for other multimedia security applications. Related work has demonstrated the usefulness of telltale watermarking for tamper recovery [12] and watermark attack characterization [15]. In [12] the authors demonstrated how telltale watermarking can be used for semiblind image restoration. In this problem the marked image undergoes unknown blurring and must be recovered using information on how the corresponding fragile watermark is distorted. In [15] the authors demonstrate how a fragile watermark can be embedded in addition to a robust watermark to characterize image tampering. The characterization process allows optimal robust watermark extraction which improves security for copyright-protection applications.

V. SIMULATION RESULTS AND COMPARISONS

A. Basis of Comparison

We evaluate the fragile watermarking techniques based on their ability to detect undesired tampering such as

- 1. Initialize user-defined variables:
 - Given: f(m, n) the host image to be watermarked.
 - Given: w(i), $i = 1, ..., N_w$, the watermark (an encrypted version of the author ID)
 - Set $L \in \mathbb{Z}^+$, the maximum wavelet decomposition level.
 - Set ckey(i), $i = 1, ..., N_w$, the coefficient selection key.
 - Set $qkey(i) \in \{0,1\}$, the quantization key.
 - Set $\delta \in \mathbb{Z}^+$, the quantization magnitude.
- 2. Perform the Lth-level discrete Haar wavelet transform on the host image f(m,n) to produce 3L detail coefficient images $f_{k,l}(m,n)$ where k=h,v,d (for horizontal, vertical or diagonal detail coefficient) and $l=1,2,\ldots,L$ is the particular detail coefficient resolution level, and a gross approximation at the lowest resolution level $f_{a,L}(m,n)$. That is,

$$\{f_{k,l}(m,n)\} := DWT_{Haar}\left[f(m,n)\right],\tag{25}$$

for k = h, v, d, a and $l = 1, \ldots L$.

3. Quantize the detail wavelet coefficients selected by ckey:

```
For l=1,2,\ldots,L, For k=h,v,d, For each (m,n), If ckey(i)=f_{k,l}(m,n) for some integer i in the range 1 to N_w, If Q_\delta(f_{k,l}(m,n))\neq w(i)\oplus qkey(i), z_{k,l}(m,n):=\begin{cases} f_{k,l}(m,n)-\delta 2^l & \text{if } f_{k,l}(m,n)>0\\ f_{k,l}(m,n)+\delta 2^l & \text{if } f_{k,l}(m,n)\leq 0 \end{cases} Else, z_{k,l}(m,n)=f_{k,l}(m,n) End End End End End End End For each (m,n), z_{a,L}(m,n)=f_{a,L}(m,n) Find
```

4. Perform the Lth-level inverse discrete Haar wavelet transform on the marked wavelet coefficients $\{z_{k,l}(m,n)\}$ to produce the marked image z(m,n). That is,

$$z(m,n) := IDWT_{Haar} [\{z_{k,l}(m,n)\}], \tag{26}$$

for k = h, v, d, a and $l = 1, \ldots L$.

Note: The function Q_{δ} is given by,

$$Q_{\delta}(f) = \begin{cases} 0 & \text{if } \left\lfloor \frac{f}{\delta 2^{l}} \right\rfloor \text{ is even} \\ 1 & \text{if } \left\lfloor \frac{f}{\delta 2^{l}} \right\rfloor \text{ is odd} \end{cases}$$
 (27)

replacement of specific image regions and their robustness to incidental image distortions such as high quality JPEG compression. In addition, we study the introduction of artifacts, if any, into the image as a result of the watermarking procedure using both qualitative observations and the peak signal-to-noise ratio (PSNR) which is defined as

$$\operatorname{SNR}(f, z) = 10 \log_{10} \left[\frac{\left(\max_{\forall (m, n)} f(m, n) \right)^{2}}{\frac{1}{N_{f}} \sum_{\forall (m, n)} (z(m, n) - f(m, n))^{2}} \right]$$
(24)

- 1. Initialize user-defined variables:
 - Given: z(m, n) the tamper-proofed image.
 - Given: w(i), $i = 1, ..., N_w$, the watermark (an encrypted version of the author ID)
 - Set $L \in \mathbb{Z}^+$, the maximum wavelet decomposition level.
 - Set ckey(i), $i = 1, ..., N_w$, the coefficient selection key.
 - Set $qkey(i) \in \{0,1\}$, the quantization key.
 - Set $\delta \in \mathbb{Z}^+$, the quantization magnitude.
 - Set T, the tamper assessment threshold.
- 2. Perform the Lth-level discrete Haar wavelet transform on the image z(m,n) to produce 3L detail coefficient images $z_{k,l}(m,n)$ where k=h,v,d (for horizontal, vertical or diagonal detail coefficient) and $l=1,2,\ldots,L$ is the particular detail coefficient resolution level, and a gross approximation at the lowest resolution level $z_{a,L}(m,n)$. That is,

$$\{z_{k,l}(m,n)\} := DWT_{Haar}[z(m,n)],$$
 (28)

for k = h, v, d, a and $l = 1, \dots L$.

3. Extract the watermark as follows:

```
For i = 1, 2, ..., N_w,
```

 $ilde{w}(i) := Q_{\delta}(z_{k,l}(m,n)) \oplus qkey(i),$

(where the coefficient $z_{k,l}(m,n)$ is selected by ckey(i), and where Q_{δ} is given by Equation 27). End

- 4. Authenticate the image:
 - (a) Decrypt \tilde{w} using the author's public key.
 - (b) Compare with the author's known identification code.
- 5. If authentication fails, assess the effect of tampering to determine level of credibility:
 - (a) Calculate $TAF(w, \tilde{w})$.
 - (b) If $TAF(w, \tilde{w}) > \mathcal{T}$

ASSESSMENT = Tampering has affected the credibility of the image.

Else

ASSESSMENT = The image is credible.

End

in decibels, where f(m,n) is the original unmarked image, z(m,n) is the tamper-proofed result, and N_f are the number of pixels in f(m,n) [or alternatively in z(m,n)] since watermarking does not increase the dimensions of the image.

We compare the performance of our technique with the watermarking methods of [4] and [5]. We do not implement the approaches in [1]–[3] for comparison as these methods provide little information to characterize the distortion and, hence, fall under a different class of techniques than our proposed algorithm.

B. Results

We demonstrate the results of the three techniques using the 256×256 image of Lena shown in Fig. 5(a). We tamper

proof the image using our proposed technique and use the following parameters: L=5 and $\delta=1.^3$ The watermark $w(i) \in \{0,1\}$ and the coefficient selection key $ckey \in \{h,v,d\}$ are randomly generated. The quantization key qkey maps the amplitude of the selected detail coefficients to binary numbers. The values of qkey are set randomly for each argument with runs of zeros and ones no greater than two to avoid visual artifacts in the marked image. We specified the qkey in this way to make the method equally sensitive to all distortions to obtain a general sense of the behavior of our technique. The resulting watermarked

³These parameters were chosen as they provide no noticeable visual change in the image. From experience, we find that $\delta=1$ is appropriate for smooth photographic images. For highly varying images, $\delta=2$ can also be used. As a rule of thumb, L may be set such that $\log_2(N/8) \le L \le \log_2(N/4)$, where N is the largest dimension of the image.





Fig. 5. Original and watermarked images for the proposed method: (a) original image of Lena and (b) watermarked image with L=5 and $\delta=1$ (PSNR = 43 dB).

Table 3The TAF Values for the Proposed Technique for Various Mean Filter Lengths

\overline{M}	l=1	l=2	l=3	l=4	l=5
3	0.5054	0.4836	0.3965	0.3281	0.2344
5	0.5035	0.4961	0.4785	0.3984	0.2969
7	0.5030	0.4934	0.4932	0.4648	0.4531
9	0.5004	0.5042	0.4561	0.4375	0.4688

image is shown in Fig. 5(b). No visual difference is noticed when viewed on a computer screen. The PSNR of the marked image is 43 dB.

As expected, TAF = 0 for the untampered marked image. If a watermark is extracted from f the unmarked image or from any other unmarked image the value of TAF is approximately 0.5. We demonstrate the effects of various image distortions such as mean filtering and JPEG compression in Tables 3 and 4. As we can see, for high-quality JPEG compression, the lower resolution subimages are still deemed credible by our method. For mean filtering, we can see from the magnitude of the TAF that the lower frequencies are less distorted than the higher frequencies. Tests were also conducted to determine

Table 4The TAF Values for the Proposed Technique for Various JPEG Compression Ratios

$\overline{\mathrm{CR}}$	l=1	l=2	l=3	l=4	l=5
2	0.3697	0.1355	0.0615	0.0000	0.0000
3	0.4977	0.2749	0.0732	0.0000	0.0000
3.5	0.4951	0.3643	0.1357	0.0703	0.0000
4.5	0.5025	0.4265	0.1455	0.0977	0.0000
5	0.4960	0.4453	0.1729	0.0469	0.0938
6	0.5052	0.4863	0.2500	0.1992	0.0469
7	0.4963	0.4824	0.3027	0.2383	0.0781
8.5	0.5029	0.5042	0.4004	0.2539	0.1094
10	0.4977	0.5059	0.4229	0.3359	0.1875





Fig. 6. (a) Tampered image. The feathers on the hat have been smoothed using an image-editing package. (b) Undistorted watermarked image.

whether localized tampering could be detected. The marked image was modified by smoothing out the feathers in the hat using an image editing package as shown in Fig. 6. The differences in the extracted watermark and embedded are shown in white in Fig. 7 for the various resolution levels. The value of the threshold to detect for tampering is application dependent. From our simulations we found that a value of approximately 0.15 allows the method to be robust to high quality compression, but detects the

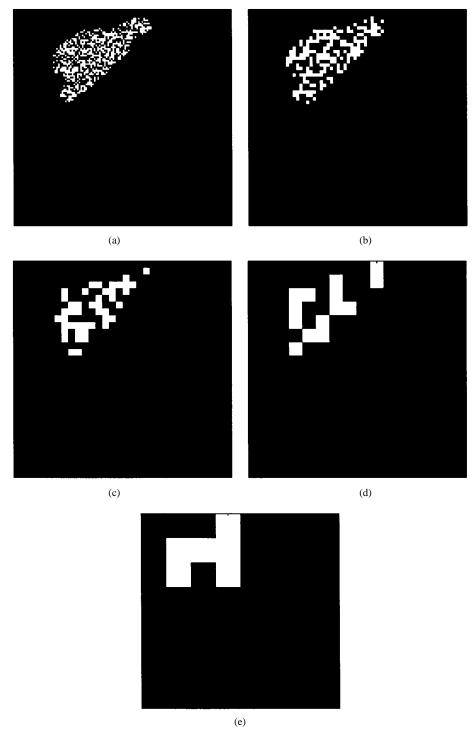


Fig. 7. Tamper detection at the various resolutions for the distorted image of Fig. 6(a) for: (a) l=1; (b) l=2; (c) l=3; (d) l=4; and (e) l=5. The differences between the embedded and extracted watermarks are shown in white for each resolution.

presence of additional tampering. A good way to analyze the effects of tampering would be to view the differences in the extracted and embedding marks as displayed in Fig. 7.

The Lena image was also tamper proofed using the method by Yeung and Mintzer [5]. The results are shown in Fig. 8. The LUT and watermark used in the technique were randomly generated as suggested. The PSNR for the marked image is 45 dB. Perfect watermark recovery

was possible when the marked image was untampered. Localized spatial regions of image tampering were also identified accurately. We tested the effects of mild filtering and JPEG compression. The results are shown in Fig. 9, where the white pixels indicate that tampering has been detected at the corresponding spatial locations. As can be seen, high-quality JPEG compression has the effect of completely destroying the credibility of the image. It is



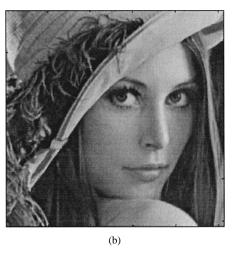
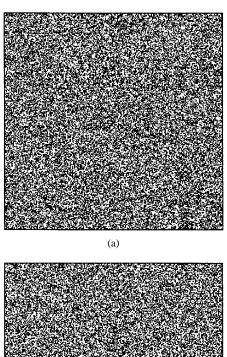


Fig. 8. Original and watermarked images for the method by Yueng and Mintzer [5]: (a) original image of Lena and (b) watermarked image (PSNR=45 dB).

not possible to distinguish between an image which is compressed with perceptual information in tact and one that is compressed in addition to being severely tampered. In general, we found that the method in [5] produced no noticeable artifacts in the marked images. The changes in the image were similar to the effect of mild dithering.

We tested the watermarking method by Wolfgang and Delp [4] on the same image of Lena. The results are shown for block sizes of 8×8 , an m sequence of order 16, and a bipolar watermark scaled by a factor of two in Fig. 10. The PSNR of the resulting image is 41 dB. As expected, no tampering was detected for the original marked image. The method detected the 8×8 blocks containing spatially localized changes (similar to that shown in Fig. 6) in the image for a threshold of zero. We found that the method identified the changes even in the presence of high quality JPEG compression (CR = 3) for a threshold set to 1.5 times the mean value of the associated autocorrelation matrix [4]. For mean filtering, the regions of high variance were detected to be tampered, but for JPEG compression, the results were more unpredictable.

To determine localized changes in the image, it is important that the block sizes be small. However, the strength of



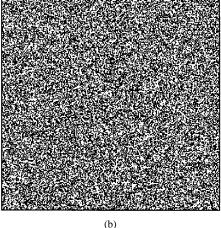


Fig. 9. Tamper identification for filtering and compression for the technique by Yueng and Mintzer [5]: (a) tamper identification for high-quality JPEG compression CR=3 and (b) tamper identification for 3×3 mean filtering.

the technique lies in statistical assumptions of the embedded m-sequence watermark, and reducing the block size lowers the statistical validity of the technique. Thus, we have found there to be a tradeoff between accuracy and localization of detection in this method. In addition, the technique requires that image-dependent information in the form of an inner product matrix [4] (which depends on the marked image and m sequence) be known for extraction. This makes the method less portable and not well suited for automation.

VI. CONCLUSIONS

Tamper proofing of multimedia signals is a new and growing field of study. Traditional approaches for data authentication are not appropriate for multimedia due to the nature of the information to be protected. In this paper we introduce the problem of telltale tamper proofing. We propose a fragile watermarking technique for images and compare its performance with existing methods. Our results indicate that the proposed approach has potential for multimedia information authentication applications. Future research involves extending this approach to characterize geometric distortions through appropriate design of the quantization key.



(a)



Fig. 10. Original and watermarked images for the method by Wolfgang and Delp [4]: (a) original image of Lena and (b) watermarked image for a block size of 8×8 , an m sequence of order 16, and a watermark scaling of a factor of two.

REFERENCES

- [1] G. L. Friedman, "The trustworthy digital camera: Restoring credibility to the photographic image," *IEEE Trans. Consumer Electron.*, vol. 39, pp. 905–910, Oct. 1993.
- [2] S. Walton, "Image authentication for a slippery new age," *Dr. Dobb's J.*, vol. 20, pp. 18–26, Apr. 1995.
 [3] M. Schneider and S.-F. Chang, "A robust content based digital
- [3] M. Schneider and S.-F. Chang, "A robust content based digital signature for image authentication," in *Proc. IEEE Int. Conf. Image Processing*, 1996, vol. 3, pp. 227–230.
- Image Processing, 1996, vol. 3, pp. 227–230.
 [4] R. B. Wolfgang and E. J. Delp, "A watermark for digital images," in Proc. IEEE Int. Conf. Image Processing, 1996, vol. 3, pp. 219–222.
- 3, pp. 219–222.
 M. M. Yeung and F. Mintzer, "An invisible watermarking technique for image verification," in *Proc. IEEE Int. Conf. Image Processing* 1997, vol. 2, pp. 680–683.
- Image Processing, 1997, vol. 2, pp. 680–683.
 [6] I. J. Cox, J. Killian, T. Leighton, and T. Shamoon, "Secure spread spectrum watermarking for multimedia," NEC Res. Inst., Princeton, NJ, Tech. Rep. 95-10, 1995.
- [7] J.-F. Delaigle, C. De Vleeschouwer, and B. Macq, "Digital watermarking," in *Proc. SPIE, Optical Security and Counterfeit Deterrence Techniques*, Feb. 1996, vol. 2659, pp. 99–110.
 [8] E. Koch and J. Zhao, "Toward robust and hidden image
- [8] E. Koch and J. Zhao, "Toward robust and hidden image copyright labeling," in *Proc. Workshop on Nonlinear Signal and Image Processing*, I. Pitas, Ed., June 1995, pp. 452–455.

- [9] G. Voyatzis and I. Pitas, "Applications of toral automorphisms in image watermarking," in *Proc. IEEE Int. Conf. Image Processing*, 1996, vol. 2, pp. 237–240.
 [10] J. J. K. Ó Ruanaidh, W. J. Dowling, and F. M. Boland,
- [10] J. J. K. Ó Ruanaidh, W. J. Dowling, and F. M. Boland, "Watermarking digital images for copyright protection," in *IEE Proc. Vision, Image and Signal Processing*, Aug. 1996, vol. 143, pp. 250–256.
- [11] D. Kundur and D. Hatzinakos, "A robust digital image water-marking method using wavelet-based fusion," in *Proc. IEEE Int. Conf. Image Processing*, 1997, vol. 1, pp. 544–547.
- [12] _____, "Semi-blind image restoration based on telltale water-marking," in *Proc. 32nd Asilomar Conf. Signals, Systems, and Computers*, 1998, pp. 933–937.
- [13] M. Abramowitz and I. A. Stegun, Eds., Handbook of Mathematical Functions with Formulas Graphs and Mathematical Tables. New York: Dover, 1965.
- [14] A. Leon-Garcia, Probability and Random Processes for Electrical Engineering. Don Mills, Ont., Canada: Addison-Wesley, 1989.
- [15] D. Kundur and D. Hatzinakos, "Improved robust watermarking through attack characterization," *Opt. Express*, vol. 3, pp. 485–490, Dec. 7, 1998.



Deepa Kundur (Student Member, IEEE) was born in Toronto, Ont., Canada. She received the Bachelor of Applied Science degree in 1993 and the Master of Applied Science degree in communications and signal processing in 1995, both from the Department of Electrical and Computer Engineering, University of Toronto, Ont., Canada. She is currently pursuing the Ph.D. degree at the University of Toronto.

Her research interests include digital watermarking of multimedia information, blind image

restoration, and data fusion for the classification of remote sensing imagery.

Ms. Kundur is currently an Engineer-in-Training with the Professional Engineers of Ontario (PEO).



Dimitrios Hatzinakos (Senior Member, IEEE) received the Diploma degree from the University of Thessaloniki, Greece, in 1983, the M.A.Sc degree from the University of Ottawa, Canada, in 1986, and the Ph.D. degree from Northeastern University, Boston, MA, in 1990, all in electrical engineering.

In September 1990, he joined the Department of Electrical and Computer Engineering, University of Toronto, where he now holds the rank of Associate Professor with tenure. His research

interests span the fields of digital signal/image processing with applications to wireless communications and multimedia. He is the author or co-author of more than 80 papers in technical journals and conference proceedings and has contributed to four books in his areas of interest. His experience includes consulting through Electrical Engineering Consociates Ltd. and contracts with United Signals and Systems, Inc., Burns and Fry Ltd., Pipetronix Ltd., Defense Research Establishment Ottawa (DREO), and Vaytek, Inc.

He has been an Associate Editor for IEEE TRANSACTIONS ON SIGNAL PROCESSING since July 1998 and Guest Editor for the Special Issue of Signal Processing on Signal Processing Technologies for Short Burst Wireless Communications, scheduled to appear in late 1999. He was a member of the IEEE Statistical Signal and Array Processing Technical Committee (SSAP) from 1992 to 1995 and Technical Program Cochair of the Fifth Workshop on Higher-Order Statistics in July 1997. He is a member of EURASIP, the Professional Engineers of Ontario (PEO), and the Technical Chamber of Greece.