Towards a Telltale Watermarking Technique for Tamper-Proofing

Deepa Kundur and Dimitrios Hatzinakos
10 King's College Road
Department of Electrical and Computer Engineering
University of Toronto
Toronto, Ontario, Canada M5S 3G4
E-mail: {deepa,dimitris}@comm.toronto.edu

Abstract

In this paper we present a novel fragile watermarking scheme for the tamper-proofing of multimedia signals. Unlike previously proposed techniques, the novel approach provides spatial and frequency domain information on how the signal is modified. We call such a technique a telltale *tamper-proofing method. Our design embeds a fragile watermark in the discrete wavelet domain of the signal by quantizing the corresponding coefficients with user-specified keys. Tamper detection is possible in the localized spatial and frequency regions of the given signal. We provide analysis, simulations and comparisons with two other tamper-proofing methods to show the potential of the proposed approach in detecting and characterizing the distortion imposed on the signal.*

1 Introduction

Research in digital watermarking has focused 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 by distorting 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 and authentication. As multimedia is often stored in digital format it is easy to modify or forge information using widely available editing software. A problem arises when the credibility of the data is under question; it must be possible to ensure that the signal has not been tampered with. We define tampering as any modification, innocent or otherwise, imposed on a given signal. Applications of this problem include authentication for courtroom evidence and insurance claims, and journalistic photography.

In this paper, we present a technique for signal tamper-proofing. Previously proposed methods [1, 2, 3, 4] 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. We embed the mark by quantizing the coefficients to a pre-specified degree using a user-defined key which provides the flexibility to make the tamper-proofing technique as sensitive to particular signal changes as desired. We call such a method a *telltale tamper-proofing scheme.*

We believe 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 are unsuitable for the estimation of a wide class of image transformations and are often costly to compute for larger images. In addition, complex models may result in an intractable framework for algorithm design. Thus, we incorporate an implicit wavelet-based model for tamper characterization.

The fundamental advantage of our technique lies in its ability to detect the resolution levels or spatial regions of the signal which are untampered and hence still credible. This is desirable when the original signal is not easily available. Existing techniques at best determine spatial regions which have undergone tampering; if high quality JPEG compression is applied to an image, previously proposed methods will assess that the entire image has undergone changes although the perceptual quality is intact. In contrast, our approach can determine frequency regions of the image which are virtually undisturbed. Therefore, an application-dependent decision can be made concerning whether

0-8186-8821-1/98 $10.00 © 1998 IEEE

the compressed image still has credibility.

2 Objectives of a Telltale Tamper-Proofing Scheme

The goal of fragile watermarking is to embed a mark in a host signal such that any changes applied to the signal will cause a change in the values of the extracted mark. Any difference between the embedded and extracted marks indicates that tampering has taken place. This is in direct contrast to robust watermarking which requires that the embedded and extracted marks be identical even when the marked signal undergoes significant distortion.

We define and discuss the objectives of a practical fragile watermarking method. To be useful such a technique must not only detect the presence of modifications in a signal, but should also provide some information helpful to characterize the distortions. The following set of criteria for an effective tamper-proofing scheme is proposed. A telltale tamper-proofing method must be able to:

1. Detect 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 additional data separate from the signal.

The attraction of a watermarking approach to tamper-proofing is that no additional data is required for signal validation. In addition, the verification information is discretely hidden in the original signal, which adds an additional level of security against attacks to modify both the signal and the verification data.

3 The Technique

The proposed method attempts to address the issues discussed in the previous section. We concentrate on the watermarking of still images. The general scenario is shown in Figure 1. A *validation key* comprised of the author's watermark, *coefficient and quantization selection keys* (which we describe later) and the quantization parameter Δ are necessary to embed and to extract the mark. The watermark can be an encrypted version of the author identification which is used to establish sender authenticity. The only user-defined parameter is the positive integer L_{max} which is the maximum wavelet decomposition level. The host image and watermark are denoted $f(m,n)$ and $w(i)$, respectively.

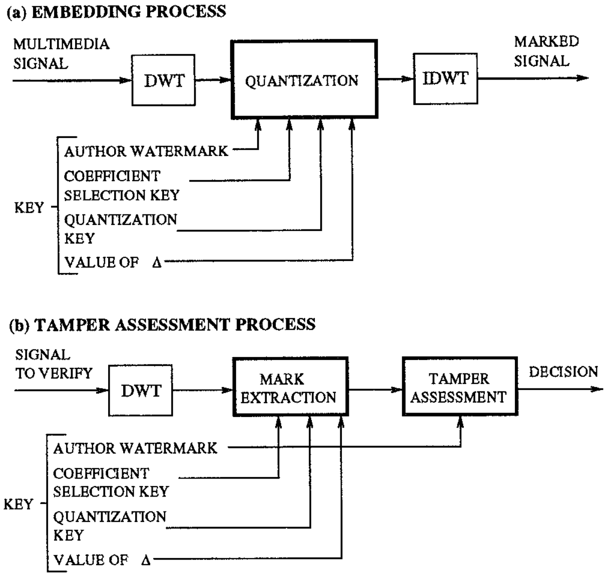

Figure 1: Proposed telltale tamper-proofing method.

The L_{max}th-level Haar wavelet transform is applied to $f(m,n)$ to produce $3L_{max}$ detail coefficient images denoted $f_{k,l}(m,n)$, where $k = h, v, d$ (for horizontal, vertical or diagonal detail coefficient) and $l = 1, 2, \ldots, L_{max}$ is the resolution level. The gross approximation at the lowest resolution level L_{max} is given by $f_{a,L_{max}}(m,n)$. The coefficient selection key $ckey(m,n) \in \{h, v, d\}$ determines the detail coefficient to mark at each location (m,n) of each resolution level. All the coefficients are not marked as we wish to minimize visual distortion.

The value of the image-dependent quantization key $qkey(m,n) \in \{0,1\}$ at each coefficient location (m,n) is a function of a localized component of the image. One purpose of this key is to make the forgery of an untampered image virtually impossible. Instead of embedding the binary watermark $w(i)$ directly into the wavelet coefficient $f_{k,l}(m,n)$, we embed $w(i) \oplus qkey(m,n)$ where $\oplus$ is the exclusive OR operator, and where $qkey(m,n)$ is dependent on the image. If we want to make the tamper-proofing especially sensitive to changes in horizontal edges of the image, then the value of $qkey$ at (m,n) and resolution l can be made a function of $f_{h,l}(m,n)$ (i.e., the values of $f_{h,l}(m,n)$ would be mapped to binary values, and $qkey(m,n)$ would represent these values). 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. Thus, the introduction of $qkey$ also provides the flexibility to monitor specific changes to the image. The algorithm is provided in Table 1.

The following function is used to mark the Haar wavelet coefficients at resolution l,

$$Q_{\Delta,l}(f) = \begin{cases} 0 & \text{if } \lfloor \frac{f}{\Delta 2^l} \rfloor \text{ is even} \\ 1 & \text{if } \lfloor \frac{f}{\Delta 2^l} \rfloor \text{ is odd} \end{cases} . \quad (3)$$

If $Q_{\Delta,l}(f_{k,l}(m,n))$ is equal to $w(i) \oplus qkey(m,n)$, then no change is made to the coefficient. Otherwise, $\pm\Delta 2^l$ is added as described in Table 1. Thus, the coefficients are quantized to pre-specified bins to reflect the bit values to embed. The addition of $\pm\Delta 2^l$ to the Haar coefficients guarantees that the signal undergoes integer changes in the spatial domain. Hence, no rounding operation is required after the IDWT operation which would be detected as tampering.

Watermark extraction is performed by taking the DWT of the potentially tampered image, and extracting the watermark bit values with the use of the validation and quantization keys. Specifically, if we let $\hat{f}_{k,l}(m,n)$ be the wavelet coefficient containing the ith watermark bit, the extracted watermark is given by,

$$\tilde{w}(i) = Q_{\Delta,l}(\hat{f}_{k,l}(m,n)) \oplus qkey(m,n). \quad (4)$$

To assess whether tampering has occurred we extract the watermark from all or some of the wavelet coefficients in the particular spatial and/or frequency regions of interest. We compute the following function which we call the *tamper assessment function* (TAF),

$$TAF(w,\tilde{w}) = \frac{1}{N_w} \sum_{i=1}^{N_w} w(i) \oplus \tilde{w}(i), \quad (5)$$

where w is the true embedded watermark, $\tilde{w}$ is the extracted mark, N_w is the length of the watermark and $\oplus$ is the exclusive OR operator. The value of $TAF(w,\tilde{w})$ ranges between 0 and 1. If $TAF(w,\tilde{w}) < \mathcal{T}$, where $0 \leq \mathcal{T} \leq 1$ is a pre-specified threshold, then the region is considered to be untampered; otherwise, tampering has occurred. The extent of tampering is given by the magnitude of $TAF(w,\tilde{w})$.

Table 1: The proposed telltale tamper-proofing technique for watermark embedding.

```
1. Perform the $L_{max}$th-level discrete Haar wavelet transform on the host image $f(m,n)$ to produce detail coefficient images $f_{k,l}(m,n)$ where $k = h, v, d$, and $l = 1, 2, \ldots, L_{max}$, and a gross approximation $f_{a,L_{max}}(m,n)$. That is,

       $\{f_{k,l}(m,n)\} := \mathrm{DWT}_{Haar}\left[f(m,n)\right],$    (1)

   for $k = h, v, d, a$ and $l = 1, \ldots L_{max}$.

2. Quantize the detail wavelet coefficients as follows:
   (a) $i := 0$.
   (b) For $l = 1, 2, \ldots, L_{max}$,
       i. $i := i + 1$.
       ii. For each $(m,n)$,
             $k := ckey(m,n)$.
             If
             $Q_{\Delta,l}(f_{k,l}(m,n)) \neq w(i) \oplus qkey(m,n)$,
               $z_{k,l}(m,n) :=$
             $\begin{cases} f_{k,l}(m,n) - \Delta 2^l & \text{if } f_{k,l}(m,n) > 0 \\ f_{k,l}(m,n) + \Delta 2^l & \text{if } f_{k,l}(m,n) \leq 0 \end{cases}$
             Else,
               $z_{k,l}(m,n) := f_{k,l}(m,n)$
             End
           End
       iii. Set $z_{a,L_{max}}(m,n) := f_{a,L_{max}}(m,n)$.
       End

3. Perform the $L_{max}$th-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,

       $z(m,n) := \mathrm{IDWT}_{Haar}\left[\{z_{k,l}(m,n)\}\right],$    (2)

   for $k = h, v, d, a$ and $l = 1, \ldots L_{max}$.
```

For authentication, the author's public key is applied to the extracted watermark to obtain the author identification code. Any tampering on the image will cause the authentication procedure to fail. In addition, the author's private key is secret so that forgery is highly improbable.

3.1 Analysis

The effectiveness of the approach to tamper-proofing is evaluated by introducing a measure we call the *tamper sensitivity function* (TSF). This is defined as the probability that tampering is detected given that $\mathcal{T} = 0$ and that N coefficients in the wavelet domain are modified. We model the degradation from image tampering on the extracted coefficients as

$$\hat{f}_{k,l}(m,n) = f_{k,l}(m,n) + w_{k,l}(m,n), \qquad (6)$$

where $f_{k,l}(m,n)$ is the undistorted wavelet coefficient, $\hat{f}_{k,l}(m,n)$ is the distorted coefficient and $w_{k,l}(m,n)$ is the associated zero mean additive Gaussian noise with variance σ^2. Two types of distortion are considered: 1) mild distortion in which we assume that $\sigma/\Delta \ll 1$; examples include high quality lossy compression (e.g. JPEG), and 2) severe distortion in which we assume that $\sigma/\Delta \gg 1$ or the effect on the extracted watermark becomes unpredictable; examples include substitution of pixel blocks and heavy filtering.

If $w_{k,l}(m,n) \neq 0$ disturbs the coefficient $f_{k,l}(m,n)$ such that the value of $Q_{\Delta,l}(\hat{f}_{k,l}(m,n))$ does not change, then tampering has not been successfully detected; we call this a false negative result. It can be shown [5] that the average probability of false negative at location (m,n) is given by

$$\overline{p}_{fn} \approx \frac{2}{\Delta}\int_0^{\Delta} \mathrm{erf}\left(\frac{\xi}{2\sigma}\right) d\xi, \qquad (7)$$

where erf$(\cdot)$ is the standard error function. Given that N wavelet coefficients are modified during the tampering procedure, it follows that

$$\begin{aligned} TSF_{mild} &= 1 - \overline{p}_{fn}^{N} & (8) \\ &\approx 1 - \left[\frac{2}{\Delta}\int_0^{\Delta} \mathrm{erf}\left(\frac{\xi}{2\sigma}\right) d\xi\right]^N & (9) \end{aligned}$$

We see that the TSF increases monotonically with decreasing Δ; hence, the smaller the value of Δ, the more sensitive the tamper detection.

Similarly, for severe distortion, it can be shown [5] that $\overline{p}_{fn} = 1/2$, and therefore

$$\begin{aligned} TSF_{severe} &= 1 - \overline{p}_{fn}^{N} & (10) \\ &\approx 1 - \left(\frac{1}{2}\right)^N. & (11) \end{aligned}$$

The TSF (in both cases) has a geometric relationship with respect to N.

4 Simulation Results

We compare the performance of our technique with the watermarking methods of Yeung and Mintzer (1997) [4] and of Wolfgang and Delp (1996) [3] using a 256 x 256 portion of the image of Lena. We tamper-proof the image using our proposed technique with $L_{max} = 5$, and $\Delta = 1$. The watermark $w \in \{0,1\}$ and the coefficient selection key $ckey \in \{h, v, d\}$ are randomly generated. The quantization key $qkey$ is generated by mapping the amplitude of the selected detail coefficients at each location (m,n) to binary numbers. The mapping from coefficient space to binary numbers is selected arbitrarily 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 behaviour of our technique. No visual difference is noticeable between the marked and unmarked images when viewed on a computer screen.

Table 2: The TAF values at each resolution level $l = 1, 2, \ldots, L_{max}$ for the proposed technique for varying mean filters of dimensions $M \times M$.

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

We demonstrate the effects of various image distortions such as mean filtering and JPEG compression in Tables 2 and 3. As we can see, for high quality JPEG compression, the lower resolution sub-images 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 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 Figure 2. The differences in the extracted watermark and embedded are shown in white in Figure 3 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 presence of additional tampering.

For the method by Yeung and Mintzer (1997) [4], localized spatial regions of image tampering were identified accurately. Tampering due to mild filtering and high quality JPEG compression was detected, but it was not possible to distinguish between an image

Table 3: The TAF values at each resolution level $l = 1, 2, \ldots, L_{max}$ for the proposed technique for various JPEG compression ratios (CR).

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
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
10	0.4977	0.5059	0.4229	0.3359	0.1875

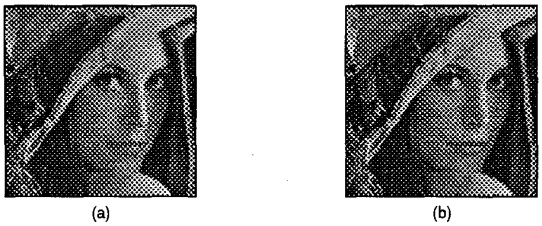

Figure 2: Example of localized spatial image tampering. (a) Marked and tampered Image. The feathers on the hat have been smoothed using an image editing package. (b) Undistorted Unmarked Image.

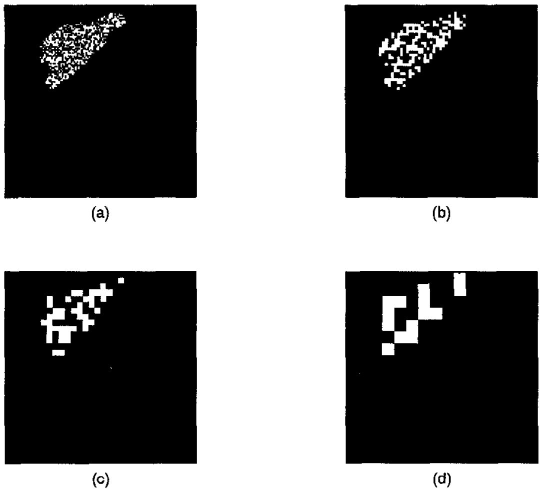

Figure 3: Tamper Detection for the Distorted Image of Figure 2 (a) for (a)$l = 1$, (b)$l = 2$, (c)$l = 3$, (d)$l = 4$. The differences between the embedded and extracted watermarks are shown in white for each resolution.

which was compressed with perceptual information completely in tact and a different unmarked image. The method by Wolfgang and Delp (1996) [3] was tested using block sizes of 8 × 8, an m-sequence of order 16 and a bipolar watermark scaled by a factor of 2. The method successfully detected the 8×8 blocks containing spatially localized changes in the image for a threshold of zero. For mean filtering the regions of high variance were detected to be tampered, but for JPEG compression, the results were more unpredictable; it was difficult to use the technique to validate the integrity of JPEG compressed images.

5 Conclusions

In this paper we introduce the problem of the telltale tamper-proofing of multimedia. The proposed fragile watermarking technique is shown to have potential for practical multimedia authentication applications. Our technique successfully identifies localized spatial and frequency regions that have undergone tampering. Our method can be used to help determine the credibility of JPEG compressed images that are perceptually identical to the original.

References

[1] S. Walton, "Image authentication for a slippery new age," *Dr. Dobb's Journal*, vol. 20, pp. 18–26, April 1995.

[2] M. Schneider and S.-F. Chang, "A robust content based digital signature for image authentication," in *Proc. IEEE Int. Conference on Image Processing*, vol. 3, pp. 227–230, 1996.

[3] R. B. Wolfgang and E. J. Delp, "A watermark for digital images," in *Proc. IEEE Int. Conference on Image Processing*, vol. 3, pp. 219–222, 1996.

[4] M. M. Yeung and F. Mintzer, "An invisible watermarking technique for image verification," in *Proc. IEEE Int. Conference on Image Processing*, vol. 2, pp. 680–683, 1997.

[5] D. Kundur and D. Hatzinakos, "Digital watermarking for telltale tamper-proofing and authentication," *submitted to Proceedings of the IEEE Special Issue on Identification and Protection of Multimedia Information*, 1998.