Semi-Blind Image Restoration Based on Telltale Watermarking

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

Abstract

We propose a novel concept for the restoration of locally degraded images based on telltale fragile watermarking. In our approach a data stream called a watermark is embedded in the wavelet domain of an image such that manipulation of the image is reflected in the embedded stream. The altered stream is used for semi-blind restoration to undo tampering. It is assumed that the embedded watermark is known, and the degradation is in the form of localized filtering in which the explicit filter characteristics are unknown. Simulation results demonstrate the potential of the approach for practical tamper recovery.

1. Introduction

The ease in which digital data can be manipulated has created a need for techniques that determine the credibility of digital information. Traditional approaches verify integrity by using additional data which is often in the form of an encrypted hashed version of the information to authenticate, but are not well-suited for the authentication and recovery of the information in digital images, sound and video. In such applications, it is desirable to be able to characterize and undo unwanted distortions [6]. Applications for tamper detection and restoration include authentication of digital data for courtroom evidence and journalistic photography.

In this paper we examine the feasibility of performing semi-blind restoration on images degraded by localized filtering. The problem of blind image restoration for global linear shift invariant (LSI) blurs has been addressed by researchers in various fields of study [2, 3, 5, 7, 9], (Also see the survey paper [4] and reference therein). To the best of the authors' knowledge this paper is the first to tackle the issue of restoration for localized image blurring when the blurring function is unknown. Our design is part of an overall image security and authentication framework based on digital watermarking. We extend our previous work on *telltale watermarking* [6] and propose a novel semi-blind restoration approach which both detects the degraded image regions and estimates elementary characteristics of the associated blurring functions.

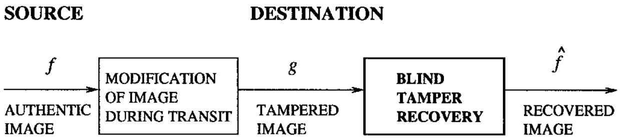

Figure 1. The image tampering and restoration scenario.

In the next section we formulate the problem we address. In Section 3 we introduce our novel approach of restoration based on telltale watermarking and analytically demonstrate how we can estimate elementary characteristics of the localized blurs to partially undo tampering. Simulation results are presented in Section 4. A discussion of the limitations of the technique and directions for future work is provided in Section 5 followed by concluding remarks in Section 6.

2. Problem Formulation

2.1. The General Tamper Recovery Problem

We address the problem of image tamper recovery. Figure 1 provides the basic scenario considered in this paper. An authentic (untampered) image f is transmitted from the source to a destination. The signal received at the destination g is modified in an unknown way. The goal of tamper recovery is to undo any modifications without use of the original image f.

We limit ourselves to situations in which the tampering is in the form of localized linear blurring. We make the following assumptions on the image degradation:

A1 $P > 0$ disjoint image regions experience localized fil-

0-7803-5148-7/98/$10.00©1998 IEEE

tering. The degraded image for each region R_p can be modeled as

$$g(m,n) = \sum_{\forall(i,j)} h_p(m-i, n-j) f(i,j), \qquad (1)$$

for $(m,n) \in R_p$, where g is the locally blurred image, f is the undistorted image, h_p is the associated blurring function for region R_p, and $p = 1, 2, \ldots, P$.

A2 The mean value of the image is preserved by the blurring. That is, $\sum_{\forall(m,n)} h_p(m,n) = 1$ for all p.

A3 The blurring function is low pass in nature. Specifically, we assume that the non-DC frequencies of the tampered image regions are attenuated by the blurring.

Localized lowpass filtering is one of the most common types of image tampering as incriminating details of the image can be easily removed and the effects are unnoticeable without access to the original image.

We do not assume that any information about the true image is available or that the explicit filter characteristics h_p are known. Thus, our recovery process attempts to solve a blind image restoration problem.

2.2. Semi-Blind Image Recovery Based on Telltale Watermarking

In all blind image restoration techniques some reference information is necessary to undo the degradation. In addition, for applications of tamper recovery we must incorporate our restoration stage within the basic image security and authentication framework. In recent studies *digital watermarking* has been proposed as an effective means to protect both the intellectual property and credibility of digital information in images [1].

In digital watermarking, a binary sequence called a *watermark* is embedded into a *host* image with the use of a user-specified *key*. The embedding process involves imperceptibly modifying the host image such that the watermark can only be extracted with the use of the key. The watermarking algorithm can be designed such that any modification of the marked image is reflected in the extracted watermark. Thus, the differences between the embedded and extracted watermarks can be used as reference for semi-blind tamper-recovery. The authors have proposed one such technique called telltale watermarking in which elementary space-frequency characteristics of the tampering can be derived [6]. In this paper, we attempt to undo the tampering by assuming that the degradation is in the form of localized filtering as suggested by Equation (1). We provide details of our technique in the next section.

3. The Proposed Approach to Semi-Blind Image Restoration

3.1. General Overview

We summarize our approach in Figure 2. The semi-blind restoration stage is a part of an overall security system in which an image f is watermarked with a binary data stream w to produce a watermarked image $\tilde{f}$. The tampered version of $\tilde{f}$ is denoted g. At the destination, the watermark is extracted to produce the binary string $\hat{w}$. If w and $\hat{w}$ are identical then it is assumed that no tampering has taken place. Otherwise, tampering is detected. The differences between the embedded and extracted watermarks are used to estimate the blurring functions and to partially restore the image.

The watermark w is embedded in the wavelet domain of the image so that the tampering can be characterized in a localized space-frequency domain. Thus, elementary characteristics of the localized blurring can be derived. Localized filtering is a form of linear shift-variant (LSV) image blurring. Since, computational complexity is of importance in our design, our approach does not try to explicitly solve the LSV image restoration problem. Instead, a multiresolution approach is employed to sub-optimally recover the image.

3.2. Telltale Watermarking

In this section we provide the relevant details of the watermark embedding and extraction stages for the restoration procedure. The reader is referred to [6] for the algorithmic specifics. The general scenario is shown in Figure 3. User-defined keys, which we do not describe in this paper, are necessary to securely embed and extract the watermark. The only user-defined parameters are the positive integers L_{max} and Δ which are called the maximum wavelet decomposition level and quantization parameter, respectively.

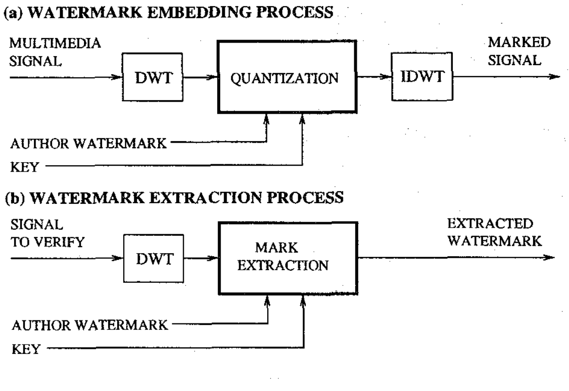

Figure 3. Proposed telltale tamper-proofing method.

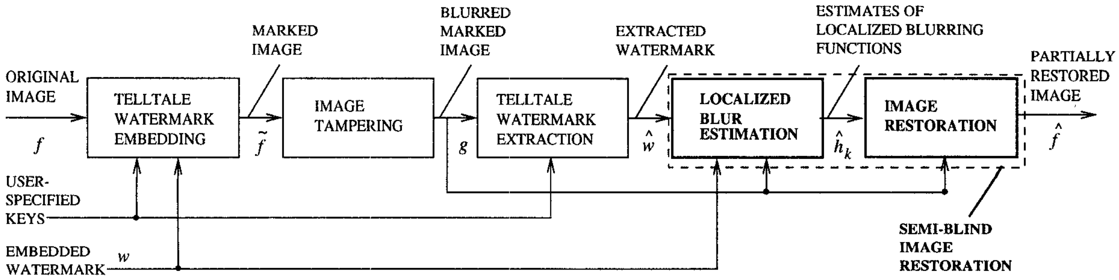

Figure 2. Proposed semi-blind image restoration technique using telltale watermarking.

To embed the watermark, the L_{max}th-level discrete wavelet transform (DWT) is applied to the original image $f(m,n)$ to produce $3L_{max}$ detail coefficient images denoted $f_{k,l}(m,n)$, where $k \in \{h,v,d\}$ (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 following rule is used to embed the ith watermark bit $w(i)$ in the coefficient $f_{k,l}(m,n))$: if $Q_{\Delta,l}(f_{k,l}(m,n))$ is equal to $w(i)$, then no change is made to the coefficient; otherwise, $-\Delta 2^l \text{sgn}(f_{k,l}(m,n))$ is added to $f_{k,l}(m,n))$ where

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

and $\text{sgn}(f) = 1$ if $f \geq 0$ and $\text{sgn}(f) = -1$ otherwise. Thus, the coefficients are quantized to pre-specified bins to reflect the watermark bit values to embed. The L_{max}th-level inverse discrete wavelet transform (IDWT) is performed on the marked coefficients to produce the watermarked image $\tilde{f}$.

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

$$\hat{w}(i) = Q_{\Delta,l}(\hat{f}_{k,l}(m,n)). \tag{3}$$

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 scaled Hamming distance between w and $\hat{w}$ which we call the *tamper assessment function* (TAF),

$$TAF(w,\hat{w}) = \frac{1}{N_w} \sum_{i=1}^{N_w} w(i) \oplus \hat{w}(i), \tag{4}$$

where w is the true embedded watermark, $\hat{w}$ is the extracted mark, N_w is the length of the watermark and $\oplus$ is the exclusive OR operator. The value of $TAF(w,\hat{w})$ ranges between 0 and 1.

Assuming that the degradation on the wavelet coefficients due to tampering can be modeled as additive Gaussian noise with variance σ_n^2, it can be shown that the expected value of the tamper assessment function for resolution level l is approximated by [6],

$$E\{TAF\} \approx 1 - \frac{1}{2^{l-1}\Delta} \int_0^{2^l \Delta} \text{erf}\left(\frac{\xi}{2\sigma_n}\right) d\xi, \tag{5}$$

where erf$(\cdot)$ is the standard error function, and it is assumed that $\sigma_n \ll 2^l \Delta$. Due to the ergodic nature of the TAF statistics in the presence of filtering, the extent of tampering at a particular resolution level l and/or spatial region can be approximated by the magnitude of the average value of the TAF at the corresponding wavelet coefficients. In the next section we show how this quantity can be related to the characteristics of the blurring functions

3.3. Analysis and Estimation of Elementary Blur Characteristics

It can be shown that the blurred image pixels in region R_p are related to the original image pixels through

$$\mathbf{g}^{(p)} = \mathbf{H}^{(p)} \mathbf{f}^{(p)}, \tag{6}$$

where $\mathbf{f}^{(p)}$ and $\mathbf{g}^{(p)}$ are the lexicographically ordered vectors containing the original image pixels and blurred image pixels, respectively, and $\mathbf{H}^{(p)}$ is th appropriate block Toeplitz blur matrix corresponding to region R_p. Furthermore, if we assume that the DWT is implemented using filtering banks and that decimation is not applied during the decomposition process, the corresponding wavelet coefficients at the lth resolution level are related by

$$\mathbf{g}_{k,l}^{(p)} = \mathbf{H}^{(p)} \mathbf{f}_{k,l}^{(p)}, \quad k \in \{h,v,d\}, \tag{7}$$

where $\mathbf{f}_{k,l}^{(p)}$ and $\mathbf{g}_{k,l}^{(p)}$ are the lexicographically ordered vectors of the DWT coefficients of the blurred image and true image, respectively, for region R_p. For the remainder of the analysis we will drop the superscript (p) for simplicity.

We model the effects of the blurring as additive noise on the embedded watermark. This "noise" at resolution l and detail component k is is given by $\mathbf{n}_{k,l} = \mathbf{f}_{k,l} - \mathbf{g}_{k,l}$. It can be shown that the autocorrelation matrix of $\mathbf{n}_{k,l}$ is

$$\mathbf{R}_{n(k,l)} = \mathbf{R}_{f(k,l)} + \mathbf{H}\mathbf{R}_{f(k,l)}\mathbf{H}^T - \mathbf{R}_{f(k,l)}\mathbf{H}^T - \mathbf{H}\mathbf{R}^T_{f(k,l)}, \tag{8}$$

where $\mathbf{R}_{f(k,l)}$ is the autocorrelation matrix of $\mathbf{f}_{k,l}$ in region R_p. We assume that the localized image region R_p can be modeled as statistically stationary which restricts the structure of $\mathbf{R}_{f(k,l)}$ to be symmetric Toeplitz. In addition, if the dimensions of the matrices are large enough (i.e., the region R_p is large compared to the spatial support of the blurring function h_p), all the matrices in Equation (8) can be successfully approximated with appropriate circulant counterparts [8]. Therefore, diagonalization can easily be carried out with the well-known Fourier matrix (which we denote $\mathbf{P}$) [8]. Thus, the power-spectral density components of the noise are approximated by the diagonal elements of

$$\mathbf{S}_{n(k,l)} = \mathbf{P}\mathbf{R}_{n(k,l)}\mathbf{P}^T, \tag{9}$$

which reduces to

$$\mathbf{S}_{n(k,l)} = \mathbf{S}_{f(k,l)} + \mathbf{S}_h\mathbf{S}_{f(k,l)}\mathbf{S}^*_h - \mathbf{S}_{f(k,l)}\mathbf{S}^*_h - \mathbf{S}_h\mathbf{S}_{f(k,l)}, \tag{10}$$

where $\mathbf{S}_{f(k,l)}$, $\mathbf{S}_{n(k,l)}$ and $\mathbf{S}_h$ are the diagonalized counterparts of $\mathbf{R}_{n(k,l)}$, $\mathbf{R}_{f(k,l)}$ and $\mathbf{H}$, respectively, and $(\cdot)^*$ is the complex conjugate matrix operator.

Assumption **A2** guarantees that the mean value of the image is preserved during blurring or equivalently that the expected value of $\mathbf{n}_{k,l}$ is zero. The variance of this zero-mean noise is given by the trace of $\mathbf{S}_{n(k,l)}$. Therefore,

$$\begin{aligned}\sigma^2_{n(k,l)} &= tr(\mathbf{S}_{n(k,l)}) \\ &= tr(\mathbf{S}_{f(k,l)}) + tr(\mathbf{S}_h\mathbf{S}_{f(k,l)}\mathbf{S}^*_h) \\ &\quad -tr(\mathbf{S}_{f(k,l)}\mathbf{S}^*_h) - tr(\mathbf{S}_h\mathbf{S}_{f(k,l)}),\end{aligned} \tag{11}$$

where $tr(\cdot)$ is the trace matrix operator and Equation (11) follows since all the associated matrices are diagonal.

We define the "elementary characteristics" of the blurring as the average attenuation at each resolution and detail component due to blurring. Because we are performing a multiresolution analysis, some frequency localization is exhibited for each l. If we approximate the bandwidth of $\mathbf{f}_{k,l}$ to be small relative to the rate of change in the values of the Fourier transform of the blur h_p, it can be shown that we can replace $\mathbf{S}_h$ in Equation (11) with $\alpha_(k,l)\mathbf{I}$ where $\alpha_(k,l)$ is a scalar constant and $\mathbf{I}$ is the appropriately dimensioned identity matrix. Thus, $\sigma^2_{n(k,l)}$ can be estimated by

$$\sigma^2_{n(k,l)} \approx \left(1 - \alpha_(k,l)\right)^2 tr(\mathbf{S}_{f(k,l)}). \tag{12}$$

Given the average value of the TAF for resolution l and detail component k, we can use Equation (5) to estimate the corresponding variance of the noise due to tampering. Equation (12) can then be used to estimate $\alpha_(k,l)$ where $tr(\mathbf{S}_{f(k,l)})$ is estimated from an untampered region near R_p. Since we assume condition **A3**, $|\alpha_(k,l)| < 1$, the solution to (12) gives a unique value for $\alpha_(k,l)$. Thus, given an estimate of $E\{TAF\}$ for a particular resolution and detail component, we can determine the average attenuation that the corresponding image details have undergone.

3.4. Semi-Blind Restoration Scenario

The following steps are used for image restoration after the watermark extraction stage shown in Figure 2.

1. Determine tampered regions R_p, $p = 1, 2, \ldots P$ [6].
2. Perform DWT on tampered image g.
3. For each R_p,
   (a) Calculate TAF for a given (k,l).
   (b) Estimate $\sigma^2_{n(k,l)}$ using Equation (5).
   (c) Estimate $\alpha_(k,l)$ with
   $$\hat{\alpha}_(k,l) = 1 - \frac{\sigma_{n(k,l)}}{\sqrt{tr(\mathbf{S}_{f(k,l)})}}. \tag{13}$$
   (d) Partially restore the image by amplifying the wavelet coefficients such that,
   If $\alpha_(k,l) < \mathcal{T}$, $\hat{\mathbf{f}}_{k,l} = \mathbf{f}_{k,l}$.
   If $\alpha_(k,l) \geq \mathcal{T}$, $\hat{\mathbf{f}}_{k,l} = 1/\alpha_(k,l)\mathbf{f}_{k,l}$.
   The threshold $\mathcal{T}$ is set to avoid noise amplification due to an ill-defined inverse problem.
4. Perform the IDWT on the restored image wavelet components $\hat{\mathbf{f}}_{k,l}$ to produce the restored image $\hat{f}$.

4. Simulation Results

We implemented and tested the algorithm of Section 3.4 to study the feasibility of our semi-blind restoration scenario. In the implementation described in [6], the Haar wavelet transform was used for its fixed point processing properties and computational simplicity; however, the inadequate frequency localization properties makes it unsuitable for semi-blind restoration. Therefore, several other wavelet transforms were considered and we provide some encouraging simulation results in Figure 4 of this section.

Figure 4(a) shows the original *host* image and Figures 4(b) and (c) show the corresponding watermarked tampered image, and restored image, respectively. The tampered region is outlined in the Figures. For the simulations we presented we used $\Delta = 2$, $\mathcal{T} = 0.2$ and the Daubechies 4 point wavelet (called "db2" in Matlab). An 11×11 radially symmetric blur given by $h(m,n) = \frac{1}{40.2}0.75^{\sqrt{m^2+n^2}}$ was applied to locally modify the image in the outlined region. The semi-blind restoration provides a peak signal-to-noise ratio improvement of 1.39 dB. In addition, it is easily seen that it improves the visual quality of the image to details.

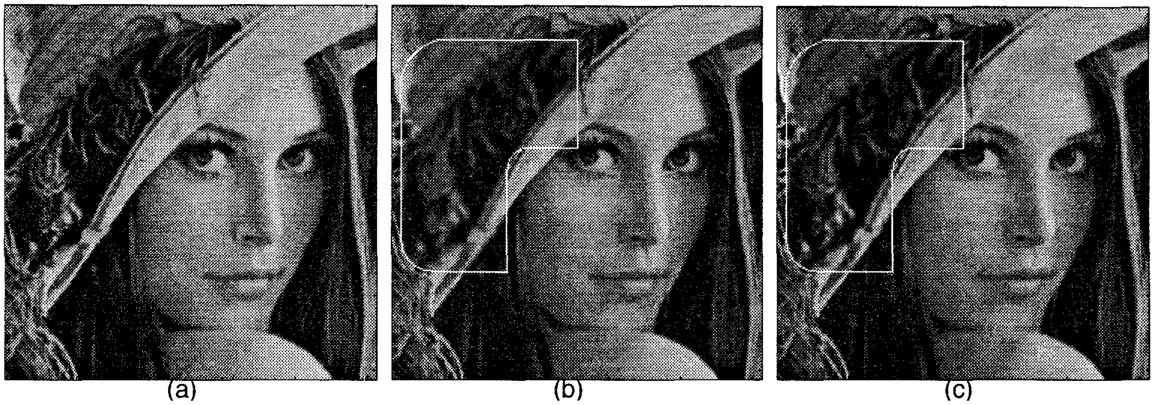

Figure 4. Simulation results. (a) Original image, (b) Tampered Image, (c) Restored Image

Several other parameters and wavelets were used in simulations. It was found that in general the larger the value of Δ, the better the restoration. This corroborates with our theoretical intuition as Equation (5) is more accurate when the relative size of Δ is large compared to the effect of the blurring. However, for large values of Δ the watermark becomes visible. In addition, larger extent Daubechies wavelets produced less watermark distortion which enables the use of larger values of Δ thus improving the restoration. Simulations also revealed that a larger tampered region significantly improves algorithm performance over smaller regions. We believe this to be because the estimates of $E\{TAF\}$ for each resolution level are more accurate.

5. Discussion and Avenues for Further Study

The fundamental limitation of the proposed approach is that the phase of the blurring cannot be determined. Further analysis not included in this paper shows that estimation of non-zero phase does not result in a unique solution. This is due to the fact that second order statistics are used in the estimation procedure. Use of third or higher order statistics can prove to result in poor statistical estimates due to the relatively small blurred data region involved in practical applications. In addition, it should be pointed out that there is an inherent trade-off in our design. The lower resolutions are more localized in frequency which results in a better piecewise constant approximation of the blurring, however, due to the poor localization in time the approximation of $E\{TAF\}$ is less accurate as fewer watermark bits are spatially embedded at these resolutions.

Future work involves investigation into wavelet transforms such as wavelet packets which could potentially give a better estimate of the localized blurring function. In addition, combining our blur estimation approach with other blind deconvolution techniques has the potential to improve overall restoration reliability.

6. Conclusions

We studied the feasibility of automatic semi-blind image restoration within a digital watermarking protection system. Improvement in the visual quality of the restored images was perceived due to the sub-optimal low complexity restoration. The performance of the algorithm is sensitive to selection of algorithm parameters, thus, further research is required to improve the reliability of the scheme.

References

[1] R. J. Anderson, I. J. Cox, S. H. Low, and N. F. Maxemchuk, editors. *IEEE Journal on Selected Areas in Communication – Copyright and Privacy Protection*, volume 16(4), May 1998.

[2] M. M. Chang, A. M. Tekalp, and A. T. Erdem. Blur identification using the bispectrum. *IEEE Transactions on Signal Processing*, 39(10):2323–2325, 1991.

[3] A. K. Katsaggelos and K.-T. Lay. Maximum likelihood blur identification and image restoration using the em algorithm. *IEEE Trans. Signal Processing*, 39:729–733, 1991.

[4] D. Kundur and D. Hatzinakos. Blind image deconvolution. *IEEE Signal Processing Magazine*, 13(3):43–64, May 1996.

[5] D. Kundur and D. Hatzinakos. A novel blind deconvolution scheme for image restoration using recursive filtering. *IEEE Transactions on Signal Processing*, 46(2):375–390, February 1998.

[6] D. Kundur and D. Hatzinakos. Towards a telltale watermarking technique for tamper-proofing. In *Proc. IEEE Int. Conf. on Image Processing*, 1998.

[7] R. G. Lane. Blind deconvolution of speckle images. *Journal aof the Optical Society of America A*, 9(9):1508–1514, 1992.

[8] H. Lutkepohl. *Handbook of Matrices*. John Wiley & Sons, Toronto, 1996.

[9] Y. Yang, N. P. Galatsanos, and H. Stark. Projection-based blind deconvolution. *Journal of the Optical Society of America A*, 11(9):2401–2409, 1994.