IMAGE AUTHENTICATION USING HIERARCHICAL SEMI-FRAGILE WATERMARKS

Yuan-Liang Tang* and Chun-Hung Chen
Department of Information Management
Chaoyang University of Technology
168, Jifong E. Rd., Wufong Township, Taichung County 41349, Taiwan (R.O.C.)

Keywords: Image authentication, Digital watermarking, Semi-fragile watermarks, Wavelet transformation.

Abstract: In this paper, a semi-fragile watermarking technique operating in the wavelet domain is proposed. A hierarchy of the image blocks is constructed and the image features are extracted such that relationships among image blocks are established in order to enhance the security and robustness of the system. With such a hierarchy, the image can be authenticated at different levels of resolution, hence providing a good property of tamper localization. In addition, by varying certain parameters, the system is able to control the degree of robustness against non-malicious attacks. The proposed algorithm thus provides a fine trade-off between security and localization, and is also robust to common image processing operations.

1 INTRODUCTION[1]

With the advance of network technologies and the popularity of digital multimedia, it is very easy to create, duplicate, transmit, and modify digital products. However, serious problems also arise along with such convenience, that is, unauthorized modification on digital products becomes very easy, too, and detection of such tampering is extremely difficult. If the digital products are images, we face the problem of *image authentication*, namely, to identify if an image has experienced malicious tampering. One of the solutions referred to as *exact authentication* embeds *fragile watermarks* (Lin, 1999) in the image and they break easily even if the image experiences only tiny modification. The applications of exact authentication are very limited because manipulations which preserve the semantics of the image should be acceptable. Such a requirement leads to another solution known as *inexact authentication*, in which *semi-fragile watermarks* (Bartolini, 2001) are embedded in stead of fragile ones. Semi-fragile watermarks are relatively robust to content-preserving manipulations, while fragile to malicious modification.

There are a number of works related to semi-fragile watermarks. For example, Queluz (1999) generated digital signatures, based on moments and edges, to protect the image. An image may be corrupted without affecting their moments, but their edges will certainly be changed. This property is used to authenticate the image content. Yu *et al.* (2000) used the Gaussian distribution to model the amount of modification on wavelet coefficients which is introduced by incidental distortions or malicious attacks. The number of coefficients necessary for watermark embedding is optimized as well. Lin *et al.* (2000) embedded a pseudorandom *m*-sequence into the median frequency DCT coefficients for image authentication. They used correlation values to determine the authenticity of an image.

[1] This research is supported by a grant from National Science Council, Taiwan, R.O.C. (NSC92-2213-E-324-024).

Tang Y. and Chen C. (2004).
IMAGE AUTHENTICATION USING HIERARCHICAL SEMI-FRAGILE WATERMARKS.
In *Proceedings of the First International Conference on E-Business and Telecommunication Networks*, pages 284-289
DOI: 10.5220/0001380902840289

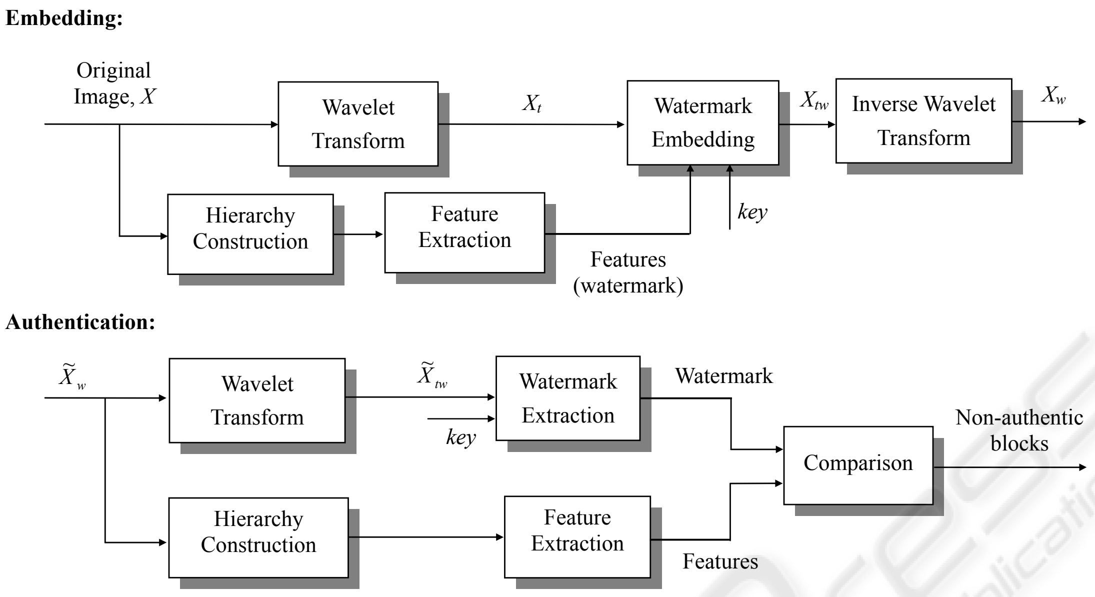

Figure 1: The embedding and authentication processes.

In addition to authenticating an image as a whole, it is also desirable to pinpoint the locations of tampering. Most localized authentication methods rely on some form of block-wise authentication (Wong, 1999), in which the image is divided into disjoint spatial regions and each of them is authenticated independently. The spatial acuity with which a block-based authentication system localizes tampering depends on the block size, thus it might be desirable to reduce the block. However, as indicated by Holliman and Memon (2000), there are potential security risks associated with smaller block sizes—the system is vulnerable to the *collage attack* (Wong, 1999; Yeung, 1997). Therefore, there exists a trade-off between security and localization. In this paper we describe a semi-fragile watermarking technique operating in the wavelet domain. A hierarchy of the image blocks is constructed and the image features are extracted such that the relationships among image blocks are established in order to resist the collage attack. The hierarchy is based on the work done by Celik *et al.* (2002), and with such a hierarchy the image can be authenticated at different levels of resolution, hence providing a good localization property. In addition, the proposed system is designed in such a way that by varying certain parameters, it is possible to control the degree of robustness against non-malicious attacks. Therefore our algorithm provides a fine trade-off between security and localization and is also robust to common image processing operations.

2 THE PROPOSED SYSTEM

Figure 1 delineates the embedding and authentication processes of our system. In the embedding process, a hierarchical structure is first constructed for the original image, *X*, and then the features are extracted from such a hierarchy. The features represent the image content and are embedded as a watermark into the wavelet-transformed image, X_t, resulting in X_{tw}. Finally, X_{tw} is inversely transformed back to the original format, producing the watermarked image X_w. The embedding process also requires a private key to ensure the security of the system. During authentication, both the image features and the watermark are extracted using the same methods as in the embedding process. These two pieces of data are then compared against each other to determine if the image blocks are authentic. The locations of tampering, if any, will be reported as well.

2.1 Construction of the Hierarchy and Feature Extraction

Celik *et al.* (2002) proposed a hierarchical structure in which the original image is first divided into disjoint blocks which constitute the bottom level of the hierarchy. And then successive levels are formed by combining distinct groups of blocks at a preceding level. Without loss of generality, we assume that 8-bit grayscale square images are dealt with. Given an $N \times N$ image X, if X is divided into $M \times M$ blocks at the bottom level, we have a hierarchy of $L = \log_2 M + 1$ levels. Let X_{ij}^l denote a block at level l, $l = 0..L–1$, where indices ij represent the spatial position. Assuming that 2×2 blocks at a given level of the hierarchy are combined to create a block at the next level, we have

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

for $l = 1..L–1$. The top level thus consists of only one block $X_{00}^{L-1} = X$. Based on Celik's hierarchy, for each block, X_{ij}^l, we first compute the mean, m_{ij}^l, of pixel intensities of the block. Due to the limitation of capacity, the bottom-level mean, m_{ij}^0, is quantized into 64 levels, i.e., a 6-bit intensity instead of the ordinary 8-bit intensity. In addition, since tampering with a block may not affect the mean when the block size is large, we introduce the *polarity* to improve the sensitivity as well as the reliability of detection. The four-bit polarity, p_{ij}^l, of X_{ij}^l is obtained by comparing the parent block's mean with those of its 4 children:

$$p_{ij}^l(x,y) = \begin{cases} 1, \text{ if } m_{ij}^l \ge m_{2i+x,2j+y}^{l-1} \\ 0, \text{otherwise} \end{cases},$$

for $l = 1..L–1$ and x, $y = 0..1$. Denoting $|\cdot|$ as the length in bits, we have $|m_{ij}^0| = 6$, $|m_{ij}^l| = 8$, and $|p_{ij}^l| = 4$ ($l = 1..L–1$), respectively. These intensity means and polarities, denoted by A_{ij}^l ($A_{ij}^0 = m_{ij}^0$ and $A_{ij}^l = m_{ij}^l \parallel p_{ij}^l$, $l = 1..L–1$), are regarded as the image features (i.e., the authentication data or watermark) and are embedded back into the image for content protection.

2.2 Watermark Embedding

The coefficients in frequency band LL_2 of the wavelet-transformed image are selected for embedding. These coefficients are good candidates in that they represent the perceptually significant part of the image and it is impossible for an attacker to tamper with the image without gross modifications to its appearance. The high level authentication data is spread over a number of lower level blocks and the accumulated payload is inserted at the lowest level of the hierarchy by wavelet coefficient modification. This is done by partitioning A_{ij}^l into a number of smaller strings:

$$A_{ij}^l = A_{ij}^l\{0,0\} \parallel A_{ij}^l\{0,1\} \parallel \ldots \parallel A_{ij}^l\{\Lambda(l)-1, \Lambda(l)-1\},$$

where $\Lambda(l) = 2^l$. The payload of a block on the lowest level is formed by concatenating the units inherited from higher level blocks:

$$\begin{aligned} D_{ij} = A_{ij}^0 &\parallel A_{C_1(i),C_1(j)}\{i - C_1(i), j - C_1(j)\} \parallel \\ &\ldots \parallel A_{C_{L-1}(i),C_{L-1}(j)}\{i - C_{L-1}(i), j - C_{L-1}(j)\}, \end{aligned}$$

where $C_b(x) = \lfloor x/2^b \rfloor$. After the above preparation, wavelet coefficients corresponding to each block on the lowest level of the hierarchy are embedded with payload bits. To increase the security level of the system, we use the pseudo-random number generator (PRNG), initialized by a private key, to establish the correspondence between an image block and the wavelet coefficients. This is illustrated in Figure 2, in which the watermark is embedded in the corresponding 4×4 coefficients in subband LL_2.

Kundur and Hatzinakos (1999) embed the watermark by first defining the quantization function:

$$Q(f,q) = \begin{cases} 1, \text{ if } kq \le f < (k+1)q \text{ for } k = 0, \pm 2, \pm 4 \ldots \\ 0, \text{ if } kq \le f < (k+1)q \text{ for } k = \pm 1, \pm 3, \pm 5 \ldots \end{cases}$$

where f is the wavelet coefficient and q denotes the size of the quantization interval. They update f by

$$f' = \begin{cases} \Delta f + 0.5q, \text{ if } Q(f,q) = b \\ \Delta f + 1.5q, \text{ if } Q(f,q) \ne b \text{ and } r > 0.5q \\ \Delta f - 0.5q, \text{ if } Q(f,q) \ne b \text{ and } r \le 0.5q \end{cases}$$

where $\Delta f = \lfloor f/q \rfloor \cdot q$, $r = f – \Delta f$ (quantization noise), and b is the watermark bit. Obviously, the result of such update will locate at exactly the middle of the quantization step, which makes it very easy to identify the watermarked coefficients. To overcome this security risk, we modify the coefficient update function as follows:

$$f' = \begin{cases} f, \text{ if } Q(f,q) = b \text{ and } (0.5q - z) \le r \le (0.5q + z) \\ \Delta f + 0.5q + s, \text{ if } Q(f,q) = b \text{ and } r > (0.5q + z) \\ \Delta f + 0.5q - s, \text{ if } Q(f,q) = b \text{ and } r < (0.5q - z) \\ \Delta f + 1.5q - s, \text{ if } Q(f,q) \ne b \text{ and } r > 0.5q \\ \Delta f - 0.5q + s, \text{ if } Q(f,q) \ne b \text{ and } r \le 0.5q \end{cases}$$

where s is a random number in the range $[1..z]$ and z is the randomness tuner ($z = \lfloor q/6 \rfloor$ in our experiments). The result of such new update will look random and therefore is more secure. Normally, a larger q gives a more robust watermark and it should vary according to the host image. However, a larger q also creates more visual impact. In order to search for an appropriate value, dozens of well-known images were tested to obtain the

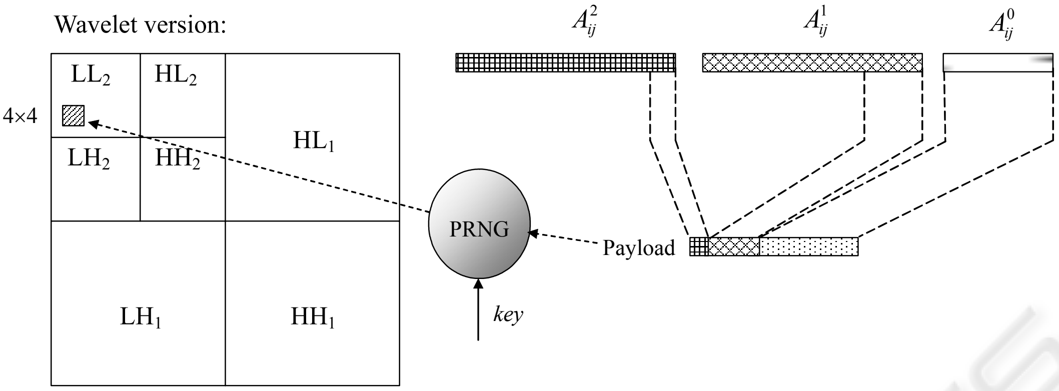

Figure 2: Concatenation of blocks to form a payload and placement of resulting payload in wavelet domain of the image.

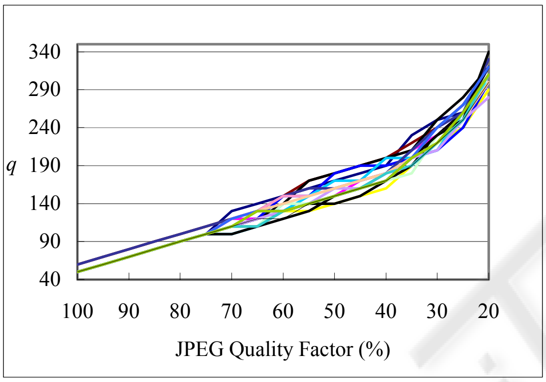

Figure 3: Watermark robustness as a relationship between quantization intervals and compression quality factors.

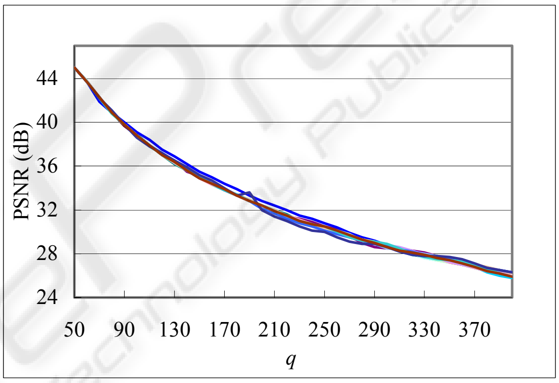

Figure 4: PSNR versus quantization intervals.

relationship between q and the JPEG quality factors (QFs). Figure 3 shows the results, in which, for a specified value of QF, any quantization intervals set below the curve will cause our system to produce false positives. For instance, if we want to validate the watermarked image that can withstand JPEG compression of 80% QF, q should be greater than 85, otherwise such an operation will be identified as a malicious attack instead of common processing. In order to obtain a more general form, those curves are approximated by a second order non-linear equation using the regression technique, and the following equation is obtained:

$$q = 393 - 6.014QF + 0.027QF^2.$$

The equation above actually defines the lower bound of the quantization interval. On the other hand, to determine upper bound, the PSNRs as a function of q are computed for each watermarked image. Figure 4 shows the results and, because PSNR ≥ 30 is generally acceptable, q should be less than 200, namely, the upper bound. Since the lower bound produces least visual impact on an image, it is a good candidate when determining q. As a consequence, our system allows determining the quantization interval automatically, depending on the visual quality requirement of JPEG compression.

2.3 Image Authentication

The authentication process is analogous to the embedding process. Let $\hat{X}_{ij}^{l}$ denote a block of the image that may have been tampered with. The watermark bit is extracted by $\hat{b} = Q(\hat{f}, q)$, where $\hat{f}$ is the corresponding coefficient. The partitioning algorithm used during embedding is reversed to recover the authentication data $\hat{A}_{ij}^{l}$, which is further partitioned to obtain $\hat{m}_{ij}^{l}$ and $\hat{p}_{ij}^{l}$. The same feature extraction is also applied to obtain $\tilde{m}_{ij}^{l}$ and $\tilde{p}_{ij}^{l}$ for each block. And finally, the difference between the extracted features and watermark is calculated. Let $T_{ij}^{l} = \left|\tilde{m}_{ij}^{l} - \hat{m}_{ij}^{l}\right|$; $\hat{X}_{ij}^{l}$ is determined as non-authentic if $T_{ij}^{l} > T^{l}$, where T^{l} is the threshold and it varies according to the size of the block. At the bottom level, since we have ignored the 2 least significant bits when collecting the authentication data, we set T^{0} = 8 (3 bits) to increase the robustness. Furthermore, because tampering with a small area may have little influence on the intensity mean of a large-sized block, the threshold should be smaller. In our experiments, we set T^{1} = 6, T^{2} = 4, and T^{l} = 2 for l = 3..L–1 to accommodate such a situation.

For polarity checks, $\hat{p}_{ij}^{l}$ and $\tilde{p}_{ij}^{l}$ are compared against each other bit by bit. Any bit difference signifies a non-authentic block. However, if the intensity means of the two blocks are similar, non-malicious modification may easily reverse their polarity. Based on such reasoning, when the intensity difference between the parent block and the child block is small, say less than 4, that bit is ignored during comparison. In summary, a block is authentic only when it passes both intensity mean and polarity tests.

3 EXPERIMENTAL RESULTS

In our experiments, the 512×512 grayscale *Lena* image is used as the host image, as shown in Figure 5(a). We set the size of the lowest level block to be 16×16 pixels, which results in a 6-level hierarchy. Figure 5(b) shows the watermarked image, whose PSNR value is about 38 dB. The degradation of the watermarked image depends on the amount of the embedded data and the embedding strength. To demonstrate the effectiveness of our technique, we modify the image by placing a tattoo (apple) on *Lena*'s arm (Figure 5(c)). As can be seen in Figure 5(d), the tampered blocks are correctly detected, in which non-authentic blocks at lower levels are shown in darker shades, while those at upper levels are shown in lighter shades. Furthermore, we perform several non-malicious manipulations to test the robustness of our system, including 80%-*QF* JPEG compression, blurring, sharpening, and addition of Gaussian noise with zero mean and variance of 20. As expected, our system didn't make any false positive errors and Table 1 shows the results.

4 CONCLUSION

We have presented in this paper an image authentication technique using semi-fragile watermarks. The authentication data is embedded in the image and is arranged in a hierarchical structure so that the whole contents of image are tightly connected in order to overcome the security weakness of block-based techniques. The system is insensitive to common image processing techniques in that robust image features are selected and a variable quantization interval further controls the degree of robustness. The system is also secure because not only the block-dependence property significantly discourages the collage attack, but also the random correspondence between blocks and coefficients prohibits brute-force attacks. The experimental results demonstrated that our system is very effective.

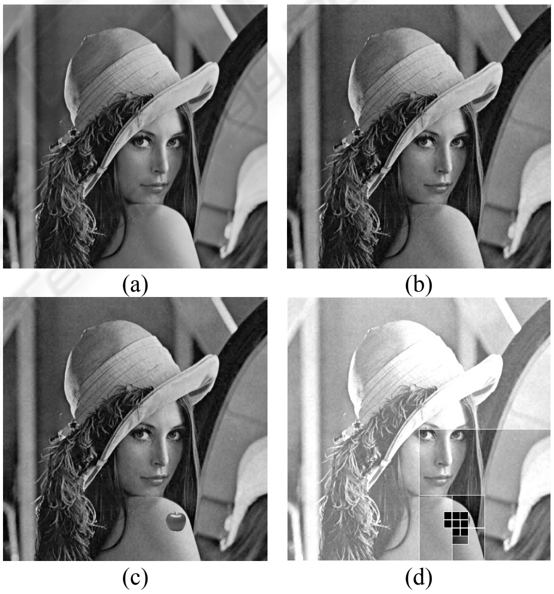
(a) (b) (c) (d)

Figure 5: (a) Original image, (b) watermarked (q = 110), (c) tampered, (d) detection result.

Table 1: Experimental results of various attacks.

Attack	Quantization Interval	Authentic?
JPEG (QF = 20%~100%)	By formula	Yes
Blurring	160	Yes
Sharpening	190	Yes
Gaussian noise addition	120	Yes

REFERENCES

Bartolini, F., Tefas, A., Barni, M., and Pitas, I. (2001) "Image authentication techniques for surveillance applications," *Proc. IEEE*, vol. 89, no. 10, pp. 1403–1418.

Bhattacharjee, S. and Kutter, M. (1998) "Compression tolerant image authentication," *Int. Conf. Image Processing*, vol. 1, pp. 435–439.

Celik, M.U., Sharma, G., Saber, E., and Tekalp, A.M. (2002), "Hierarchical watermarking for secure image authentication with localization," *IEEE Trans. Image Processing*, vol. 11, no. 6, pp. 585–595.

Chotikakamthorn, N. and Sangiamkun, W. (2001) "Digital watermarking technique for image authentication by neighbouring block similarity measure," *Int. Conf. Electrical and Electronic Technology*, vol. 2, pp. 743–747.

El-Din, S.N. and Moniri, M. (2002) "Fragile and semi-fragile image authentication based on image self-similarity," *Int. Conf. Image Processing*, vol. 2, pp. 897–900.

Fridrich, J., Goljan, M., and Du, R. (2001) "Invertible authentication watermark for JPEG images," *Int. Conf. Coding and Computing*, pp. 223–227.

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

Kundur, D. and Hatzinakos, D. (1999) "Digital watermarking for telltale tamper proofing and authentication," *Proceedings of the IEEE*, vol. 87, pp. 1167–1180.

Lin, E.T. and Delp, E.J. (1999) "A review of fragile image watermarks," *Multimedia and Security Workshop* (*ACM Multimedia* '99) *Multimedia Contents*, pp. 25–29.

Lin, E.T., Podilchuk, C.I., and Delp, E.J. (2000) "Detection of image alteration using semi-fragile watermarks. *SPIE Conf. Security and Watermarking of Multimedia Content II*, vol. 3971, pp. 152–163.

Queluz, M.P. (1999) "Content-based integrity protection of digital images," *SPIE Conf. Security and Watermarking of Multimedia Contents*, vol. 3657, pp. 88–93.

Sun, Q. and Chang, S.-F. (2002) "Semi-fragile image authentication using generic wavelet domain features and ECC," *Int. Conf. Image Processing*, vol. 2, pp. 901–904.

Wong, P.W. (1999) "A watermark for image integrity and ownership verification," *IS&T Image Processing, Image Quality, Image Capture, Systems Conference*, pp. 374–379.

Yeung, M. and Mintzer, F. (1997) "An invisible watermarking technique for image verification," *Int. Conf. Image Processing*, vol. 2, pp.680–683.

Yu, G.J., Lu, C.S., Liao, H.Y. Mark, and Sheu, J.P. (2000) "Mean quantization blind watermarking for image authentication," *Int. Conf. Image Processing*, vol. 3, pp. 706–709.