

Content based Public Key Watermarking Scheme for Image Verification and Authentication

Jayashree S. Pillai
Associate Professor
AIMS, Peenya, Bangalore

Annamma Abraham, Ph. D
Professor
BMS Institute of Technology, Bangalore

ABSTRACT

A semi fragile, public key based digital signature scheme to verify and authenticate digital images is proposed. The scheme computes a content dependent feature value for each block of DCT coefficients of the original image. This feature is signed and then embedded into the selected DCT coefficients of the image as watermark. The technique aims to circumvent the cut and paste attacks by incorporating inter block dependence during the watermarking procedure. The watermarked image demonstrated high perceptual quality and robustness. It also proved to be robust against many common image processing operations and JPEG compression. The scheme not only can verify the authenticity and the integrity of images, but also can locate the illegal modifications.

General Terms

Image watermarking, authentication, content based watermarking

Keywords

Inter block dependence, digital signature, content authentication, semi fragile watermarking, block wise percentage difference

1. INTRODUCTION

Image authentication [1][2][3] can mainly be considered as the procedure to 1) verify the source/owner of the image – source authentication and 2) ensure that the content of the image has not been modified – integrity of the image. The need for image authentication and integrity is well emphasized in several imaging applications including legal use of images, archiving and sharing of medical images, copyright in news broadcasting systems and commercial transactions. In all these applications, there is an inherent requirement to verify the integrity and authenticity of the media. Image authentication can be done using a number of approaches like conventional cryptography, fragile and semi fragile techniques and content based signatures.

Digital watermarking [4][5][6][7][8] is a widely studied procedure to achieve image authentication. Digital signatures and traditional encryption techniques have been used but watermarking permits the authenticating information to be embedded into the image by exploiting various properties of the image like color, edges, luminance, signature and statistical features like mean and standard deviation. The image is considered as a communication channel for the watermark embedded as a noise pattern. The embedded watermark, usually imperceptible, may contain either a specific owner ID or some content-related features that can be used for authentication.

The authentication technique should be useful to identify the source of the image and alterations that occur after the insertion of the signature or watermark.

Watermarking techniques can be classified as fragile[8][9], semi fragile[10][7][11] and robust [12][13] based on the requirements of the authenticator. Fragile watermarks are easily destroyed by the slightest modification to the image where as robust watermarks are highly tolerant to most of the operations to manipulate the watermark or image that at times it is not able to differentiate between the intentional tampering and common image processing operations like jpeg compression or mild noise. Semi fragile watermarks are a compromise between the two. It is very common for an image to be compressed to save bandwidth and space and filtered to improve the image quality. The transmissions of images over noisy channels introduce noises into the image. The semantics of the image is not affected here. The distortions are considered incidental and an image authentication system should be able to tolerate such mild distortions and at the same time should be able to ring alarm bells whenever an intentional modification or tampering operation is encountered that is aimed at modifying the content or meaning of the content.

In this paper, a semi fragile watermarking technique is proposed to improve on Chang's scheme[14] to make it intolerant to jpeg compression and other common signal processing operations. An image feature generated from the DCT of the image is used as a watermark. To thwart cut and paste attack, as indicated by Memon and Holliman[15], inter block dependence is considered and the watermark is inserted in the transform domain.

The rest of the paper is organized as follows. Section II briefs the related work in the areas of block wise public key watermarking and content based watermarking, Section III explain the proposed algorithm, Section IV briefs the experimental results and Section V concludes the work.

2. RELATED WORK

2.1 Wong's Scheme

Wong[16][1] proposed a public key fragile marking technique in the spatial domain for image integrity verification. The original image is partitioned into blocks and a digest is generated for each block using a hash function, the inputs to which include the MSB's of the block with the LSB set to zeros and the dimensions of the block, the image id and the block number. First few bits of the output of the hash function, the size of the block, is then element wise exclusively ORed with the watermark image and then encrypted with the private key of the authenticator to get the

signature. The signature is embedded to the least-significant bit plane of the original image.

The detector or image authenticator then extracts the signature from the LSB of the received image using the authenticator's public key. The signature is also regenerated from the MSB's of the received image. The authentication is done using correlation of the extracted and derived signatures. This technique has localization properties and can identify regions of modified pixels within a marked image as any change to the image will result in changes to the binary watermark. The main goal of this scheme is to ensure the image's ownership and integrity. The issue is that it is susceptible to counterfeiting and transplantation type of attacks as the blocks are independent. The watermark is embedded in the spatial domain and is intolerant to jpeg compression and other incidental image processing operations like filtering and noise.

2.1 Baretto et.al's Scheme

To solve the issue of counterfeiting type of attacks in Wong's scheme, [17][18] adopted the strategy of hash block chaining – HBC1, which introduces dependencies on the neighboring blocks and HBC 2, which makes use of non-deterministic digital signature to achieve inter block dependency among the image blocks. The feature value for each block is not dependent only on that block but also on the subsequent block. The domain of embedding is spatial and intolerant to jpeg compression and other incidental image processing operations like filtering and noise.

2.2 Chang et.al's scheme

Chang[14][19] used the concept of continued fractions to make the blocks interdependent. Statistical feature like mean was used as the content based feature to determine the watermark. It is an extension of Wong's scheme [8] and aims to achieve inter block dependence to thwart the attacks cut and paste attack.

The image is broken down into sub blocks, the LSB for each sub block is set to zero and a hash is calculated using the mean of the MSB's, the continued fraction value for the block, the block dimensions and the signature value of the previous block. The hash output is signed using the owner's private key to generate the signature of that block. This signature is embedded into the LSB of the block. The verification is done by extracting the signature from the LSB of the received image, decrypting it with the owner's public key and correlating it with the hash value calculated from the MSBs. The watermark is embedded in the spatial domain and is intolerant to the incidental attacks like JPEG compression, filtering and noise attacks.

3 THE PROPOSED SCHEME

Any content authentication system, in general, consists mainly of four parts. In this section the algorithm that extracts the feature value and embeds the signature calculated using the feature value into the selected DC coefficients is explained. The algorithm is intended to enhance Chang's algorithm to make it intolerant to incidental image processing operations like JPEG compression, filtering and Gaussian noise.

The image is segmented into blocks of size 8 x 8 and DCT coefficients are derived. The mean of 10 low frequency DCT coefficients of block is computed. The feature value of the block is computed from the image content as a real number A/B where A is the mean value of the current block and B is

the mean value of the subsequent block. This feature value is one of the components used as input to the hash function which is then encrypted using the private key to generate the unique watermark for that block. The feature also serves as a chain to achieve inter block dependency. This inter block dependency insulates against the cut-paste and transplantation attack which is one of the most common modification attack.

The watermark generation, embedding, extraction and verification procedures are as follows:

3.1 Watermark generation

1. Segment the host image X into n blocks and X_r represents each block
2. Calculate DCT D_r of each block X_r
3. Calculate the mean M_r of ten low frequency coefficients of each block including the DC component as in Figure 1
4. Calculate the feature value for each block

$$F_r = M_r / M_{r+1}, \quad 0 \leq r \leq n-1$$

where M_r is the mean value of the current block and M_{r+1} is the mean value of the subsequent block. The calculation of F_r for a block follows the below mentioned pattern, given in Figure 1, and for the last sub block, i.e., $r = n-1$,

$$F_{n-1} = M_{n-1} / M_1$$

5. Steps 1 to 4 is repeated for all the blocks

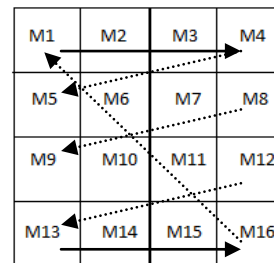


Figure 1: Zig zag pattern of inter block dependence

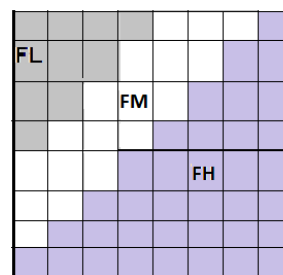


Figure 2: Frequency regions in 8 x 8 DCT. FL represents selected lower frequencies

3.2 Watermark embedding

1. Select the mid frequency coefficient at the selected location (p,q) in each DCT block
2. Replace the chosen coefficient with the signed watermark

$$DCT(p, q) = \text{sign}(DCT(p, q)) * (\alpha * w_i)$$

where α is the calculated value for embedding strength and the sign of the coefficient is retained

3. Repeat steps 1 and 2 for all the blocks
4. Perform inverse DCT to get the watermarked image X^*

3.3 Watermark extraction procedure

1. Segment the received watermarked image X^* into n blocks and X_r^* represents each block
2. Calculate DCT D_r^* of each block X_r^*
3. Extract the embedded watermark from the mid frequency coefficient of each block

$$w_i = |DCT(p, q)| / \alpha$$

and compute the value of the mid frequency coefficient after extraction of the watermark to get the image I

4. Repeat step 3 to get the extracted watermark

$$W^* = \{w_1^*, w_2^*, \dots, w_n^*\}$$

3.4 Watermark verification procedure

1. Re-compute the watermark on the image I using steps 3 – 6 of the watermark generation procedure to obtain the computed watermark

$$W^{\sim} = \{w_1^{\sim}, w_2^{\sim}, \dots, w_n^{\sim}\}$$

2. Perform inverse DCT of the image I
3. Calculate the block wise percentage difference

$$\Delta = |w_i^* - w_i^{\sim}| / \max\{w_i\} * 100$$

The block wise percentage difference Δ is dependent on the image and type of application and authentication requirement. It is used to detect any manipulation or changes in the block. If the percentage difference is smaller than the experimentally chosen value, the block is considered to be authentic. Otherwise that block is considered to be tampered with.

4 EXPERIMENTAL RESULTS

The proposed scheme was evaluated using various grey scale images. The images were subject to jpeg compression, filtering and incidental noise and were also manipulated using Photoshop to demonstrate image modification attacks. Images of different formats like tiff, bmp and png were used in the experiments.

4.1 Selection of parameters for the experiment

4.1.1 Block size

To study the algorithm, the test images were segmented into blocks of size 8 and 16. From the experiments, it was concluded that block size of 16 ensures efficiency in time often resulted in a better PSNR value. The robustness to JPEG attack is better when block size of 8 is used.

4.1.2 Feature value

Statistical feature mean is used to represent the feature and watermark of the image block. The DC coefficient is the most important component of the block and represents the average value of the block. The mean is calculated for the DC coefficient and nine other low frequency coefficients. Any incidental operation like compression and filtering normally affects the high frequency coefficients and the choice of the mean of low frequency coefficients is to ensure that the feature value is not affected by the incidental frequency dependent operations.

4.1.3 Mid frequency coefficients

In order to embed the watermark in the transform domain, embedding in the low frequency coefficients resulted in visual degradation of the image and embedding in the high frequency coefficients is normally not advisable as it may be lost during compression. Any mid frequency or combination of mid frequency coefficients could be used to embed the watermark. In this experiment, the watermark was embedded in the mid frequency coefficient of each block given by middle element of the mid diagonal ($\text{blocksize}/2, \text{blocksize}/2$)

4.1.4 Embedding strength α

This value is used to scale the watermark appropriately so that it has the same range as that of the DCT coefficients. In order to determine the suitable value, statistics like the mean, median or standard deviation is generally used. In this experiment, the standard deviation α_x of the DCT coefficient values at the mid diagonal of all the blocks of the image is obtained. The embedding strength α is calculated as α_x / α_w , where α_x is the standard deviation of the mid frequency coefficient values and α_w is the standard deviation of the computed value of watermark for each block.

4.2 Results

4.2.1 Quality of the watermarked image

The quality metrics used to measure the quality of the watermarked image is PSNR – Peak Signal to Noise Ratio, PCC – Pearson Correlation Coefficient, Image quality index and SSIM – structural similarity. The metrics are calculated for the original and watermarked image.

The tests results as in Table 1 demonstrate a very high level of similarity between the original and watermarked image. In this algorithm, the average PSNR value is around 54 which

indicates high quality watermarked image. The average values of PCC, image quality index and SSIM is very close to 1 which indicates that the watermark is imperceptible. The histograms of the original and watermarked LENA images as in Figure 3 are also similar.

4.2.1 Tolerance to incidental operations

It was observed that the watermarked image is quite tolerant to normal image processing operations like JPEG compression at various quality levels, Gaussian and salt and pepper noise and median filtering for grey scale images. It was not as robust against histogram equalization as during this processing, the pixel values change considerably as in table 2

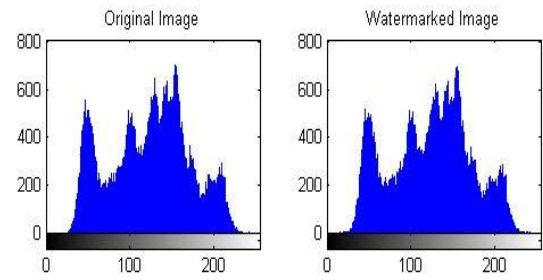


Figure 3: Histogram of the original and watermarked LENA 256 image

4.2.2 Robustness

The robustness of the watermarking technique is quite high where 8 bits per block is embedded. In a 256x256 image, using block size 16, 4096 bits are embedded and in 512x512 images, 16384 bits are embedded s watermark.

The time to embed, extract and authenticate the watermark is around 4sec for a 256x256 pixel image and around 12sec for 512x512 size image on a Pentium system.

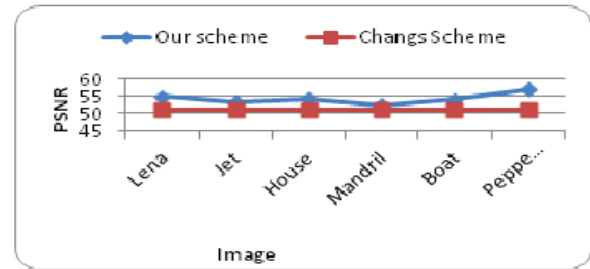


Figure 4: Comparison of PSNR values of proposed algorithm with Chang's scheme

Chang's embedding procedure showed an average PSNR value of 51.12 for various images. Our algorithm showed a better PSNR value for the same set of images indicating the quality of the embedded watermark is better.

4.2.3 Detection of tampering of watermarked image

The original images were subject to intentional attacks like cut and paste attacks and modification attacks using Adobe Photoshop, Paint and other image processing tools. The watermark extraction algorithm could identify the regions of attack Figure 6 and appropriately report the blocks that were tampered with.


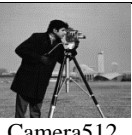

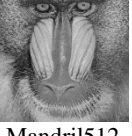
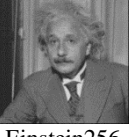
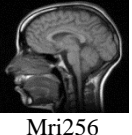
Image after watermarking	PSNR	PCC	Image quality index	SSIM
 Lena256	55	0.99	0.98	0.99
 Camera512	53.5	0.99	0.99	0.99
 livingroom512	54.3	0.99	0.99	0.99
 Mandril512	52.5	0.99	0.99	0.99
 Einstein256	54.1	0.99	0.99	0.99
 Mri256	60.2	0.99	0.98	0.99
Average	54.9	0.99	0.98	0.99

Table 2. Results of watermark extraction after incidental signal processing operations of cameraman512 image

Attack	Quality Parameter	Highest % Difference
Jpeg Compression		
Maximum	100%	0.01
High	90%	0.64
Medium	80%	0.71
Average	70%	3.2
Gaussian Noise	M=0 V=0.01	4.38
Salt & Pepper Noise	0.01	7.25
	0.02	9.3
Histogram equalization	0-200	60.4
Median Filter	3x3	8.32

Dimension of the host image are a part of the watermark and any change in the image size will result in an output that resembles random noise as the change in dimension will affect every block of the image.

**Figure 5: Living room images after compression using 90%, 80% and 70% quality parameters**

When the wrong public key is used during the verification procedure, all the blocks are reported as unauthentic.

**Figure 6 - Tampering attacks on living room (512x512) and Lena (256x256) images**

For the block wise percentage difference Δ , a threshold value of less than 15 indicated that the distortions were incidental

and a value > 15 considered the block to be intentionally tampered with. Details in Table 3.

Table 3. Detection of tampering of watermarked images

Image	Attack	Blocks identified as tampered
Living room	Vase on the right side removed	945, 946, 977 and 978
Lena	Image inserted in right bottom corner	223, 224, 240, 254, 255, 256

5 CONCLUSION

This paper presents a semi fragile watermarking technique using image content as the watermark. No additional watermark is needed. The ratio of mean values of the significant DCT coefficients of each block is used as the feature value and is embedded in the mid frequency coefficients of each image block. The test results indicate the watermarking algorithm is robust to incidental image processing operations like JPEG compressions and other as indicated in Table 2 and is still able to authenticate it.

The quality of the watermarked images is better than Chang's scheme as in Figure 4. The algorithm is able to identify tampered portions of the image with accuracy. Moreover, the algorithm circumvents the cut and paste attack as the blocks are dependent based on the chosen pattern and for every block that is modified, at least one more block is detected as not authentic. The algorithm does not recover the blocks that have been identified as tampered and future work will be directed towards recovery of portions of tampered blocks.

6 REFERENCES

- [1] P. W. Wong and N. Memon, "Secret and public key image watermarking schemes for image authentication and ownership verification," *IEEE Trans. Image Process.*, vol. 10, no. 10, pp. 1593–601, Jan. 2001.
- [2] N. M. Wong, P. W., "Secret & Public authentication WM schemes that resist VQ attacks."
- [3] L. Yang, J. Tian, and D. Wu, "Content-based image authentication by feature point clustering and matching," *Secur. Commun. Networks*, vol. 5, no. 6, pp. 636–647, Jun. 2012.
- [4] I. J. Cox and M. L. Miller, "The First 50 Years of Electronic Watermarking," *EURASIP J. Appl. Signal Processing*, pp. 126–132, 2002.
- [5] M. L. Miller, I. J. Cox, and J. M. G. Linnartz, "A review of watermarking principles and practices," *Publ. "Digital Signal Process. Multimed. Syst. Ed. K. K. Parhi T. Nishitani, Marcell Dekker Inc."*, no. February 1997, pp. 461–485, 1999.
- [6] R. J. Anderson and F. A. P. Petitcolas, *INFORMATION HIDING AN ANNOTATED BIBLIOGRAPHY*. 1999, pp. 1–62.
- [7] A. Haouzia and R. Noumeir, "Methods for image authentication: a survey," *Multimed. Tools Appl. # Springer Sci. + Bus. Media, LLC* 2007, vol. 39, no. 1, pp. 1–46, Aug. 2007.
- [8] E. T. Lin and E. J. Delp, "A Review of Fragile Image Watermarks," *CERIAS Tech Rep. 2001-74*, 2001.

- [9] C. S. Engineering, P. Jain, and A. S. Rajawat, "Fragile Watermarking for Image Authentication : Survey," *Int. J. Electron. Comput. Sci. Eng.*, 1956.
- [10] E. T. Lin, C. I. Podilchuk, E. J. Delp, and M. Hill, "Detection of image alterations using semi-fragile watermarks."
- [11] G. K. Ci, E. Janu, K. Cius, and H. Schumann, "Tamper-Proof Image Watermarking , Based on Existing Public Key Infrastructure," vol. 16, no. 1, pp. 1–18, 2005.
- [12] K. Hung, L. Shiang, and Y. Road, "A Novel Robust Watermarking Technique Using IntDCT Based AC Prediction," vol. 7, no. 1, pp. 16–24, 2008.
- [13] C. Lin and S. Chang, "Generating Robust Digital Signature for Image / Video Authentication," no. September, 1998.
- [14] C.-C. C. and W.-C. Wu, "Public-Key Inter-Block Dependence Fragile Watermarking for Image Authentication Using Continued Fraction," *Informatica*, vol. 28, no. 2, pp. 147–152, 2004.
- [15] M. Holliman and N. Memon, "Counterfeiting Attacks on Oblivious Block-wise Independent Invisible Watermarking Schemes," *IEEE Trans. IMAGE Process.*, vol. 9, no. 3, pp. 432–441, 2000.
- [16] P. W. Wong and W. Road, "A Public Key Watermark for Image Verification and Authentication," pp. 455–459, 1998.
- [17] hae Y. K. s l m Barreto, "Pitfalls in public key watermarking."
- [18] P. S. L. M. Barreto, H. Y. Kim, and V. Rijmen, "Toward secure public-key blockwise fragile authentication watermarking," *IEE Proc. - Vision, Image, Signal Process.*, vol. 149, no. 2, p. 57, 2002.
- [19] W. Tai, C. Yeh, and C. Chang, "Reversible Data Hiding Based on Histogram Modification," *IEEE Trans. CIRCUITS Syst. VIDEO Technol.*, vol. 19, no. 6, pp. 906–910, 2009.