ON THE SECURITY OF STRUCTURAL INFORMATION EXTRACTION/EMBEDDING FOR IMAGES

Chun-Shien Lu

Institute of Information Science, Academia Sinica,
Taipei City, Taiwan 115, ROC
e-mail: lcs@iis.sinica.edu.tw

ABSTRACT

In addition to robustness and fragility, security is a quite important issue in media authentication systems. This paper first examines the insecurity of several block-based authentication methods under counterfeit attacks. Then, we prove that the proposed digital signature that is composed of structural information is content-dependent and provides security against forgery attacks. Experimental results demonstrate the benefits of exploiting structural information in a media authentication system.

1. INTRODUCTION

1.1. Image Authentication

Methods of multimedia content authentication can be categorized into either digital signature-based or watermarking-based. Digital signature (or robust hash) is basically a condensed representation or perceptual digest, which captures the essence of a media content. It is stored as an extra file and later used for authentication. Watermarking, on the other hand, is an invasive method that really embeds a message into a media data and the hidden message is later extracted to verify the authenticity of a media content.

In [11], Wong proposed a block-wise fragile image watermarking method to detect tampered areas. At the watermark embedding stage, an image is first divided into non-overlapping blocks and each block is watermarked individually. In [5], Lin and Chang proposed a block DCT-based robust digital signature method for image authentication. Digital signature bits are quantized results, which are generated from comparing selected pairs of DCT coefficients in disjoint block pairs determined by a secret key.

1.2. Security vs. Counterfeit Attacks

Counterfeit attacks [2, 4] were addressed to raise the insecurity of [5, 11]. Security means the ability of deterring attackers from being able to forge an arbitrary image that can be authentic. Traditional block-based methods [5, 11] are common in that they considered disjoint blocks where no contextual information exists between them. With this prior knowledge, Holliman and Memon [4] proposed a counterfeit attack such that a forged image can still be authentic. Assume that attackers know the dimensions of a block unit and the binary logo but don't know the secret key used for hashing. The goal is to forge an arbitrary image Y^* from a set ($\mathcal{A}$) of available authentic images such that it is perceptually similar to Y ($\notin \mathcal{A}$) and Y^* can be detected to contain watermarks that were previously inserted in the each element of $\mathcal{A}$. At first, the image Y is divided into blocks from which block search is stirred in $\mathcal{A}$. Only the best block match (in terms of minimum mean square error) is used to constitute the block unit of Y^*. The above process is repeatedly performed. In addition, Fridrich *et al.* [2] proposed a so-called "collage attack" to further improve the perceptual quality of a forged image generated from [4] by error diffusion.

As for [5], Radhakrisnan and Memon [10] proposed a counterfeit attack to address an insecurity issue. Based on the assumption that a set of authentic images is available and the same key is used to generate mapping of block pairs, their attack is to deduce one block pair at each time according to an incoming signature bit. By repeatedly executing the above process, the mapping function of block pairs can be deduced. Thus, the corresponding pairs of DCT coefficients could be modified such that the desired magnitude relationship could be purposely created.

To maintain the security of watermarking-based authentication systems while not sacrificing the localization accuracy and increasing the complexity, content-dependent key [3] was addressed. The major characteristic is that the content-dependent key has to be extracted from a media content itself at the watermark generation and verification stages, respectively. However, it is not guaranteed to always produce the same key when media content has been distorted (This corresponds to a robustness issue.).

The underlying philosophy of [5, 11] is their independent block-based characteristic. The absence of employing structural information leads to the risk of counterfeit attacks. In this paper, we shall describe how a piece of structural in-

formation could be used to overcome the aforementioned counterfeit attacks. In addition, our content-dependent digital signature contains a kind of structural information that can be publicly known. Several results are provided to demonstrate the security of structural digital signature.

2. EXTRACTION OF STRUCTURAL INFORMATION AS DIGITAL SIGNATURE OR WATERMARK

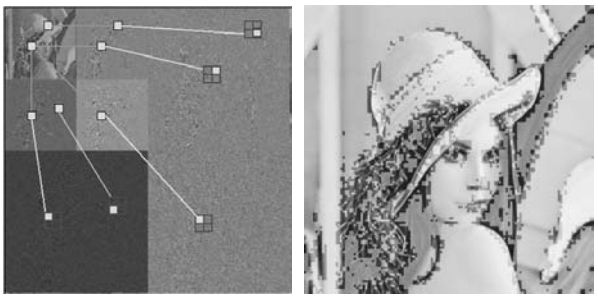
We explore the security issue of the structural digital signature (SDS) scheme [7]. The SDS is constructed in the wavelet transform, which offers multiscale space-frequency localization and allows to design a digital signature with structural but secure information. In the wavelet domain of an image, the so-called joint (interscale) parent-child pairs exist. Parent-child pairs have been confirmed to be uncorrelated but statistically dependent [1]. This dependency mainly arises from the perceptually important semantic features, e.g., edges and textures. Based on these semantic features, the so-called structural digital signature is constructed to simultaneously resist against incidental manipulations (e.g., JPEG/JPEG2000 compression) and reflect malicious distortions. The construction of an SDS is summarized in Table 1. Fig. 1 illustrates some selected parent-child pairs in the wavelet domain and the result mapping back into the spatial domain. Each parent-child pair represents a magnitude relationship involving different resolutions (frequencies). We will describe in the next section that this relationship is extremely difficult to forge. The extracted SDS has been stored for image authentication [7] or embedded for error concealment [6] and anti-disclosure of watermarks [8]. The main theme of this paper will be investigating the security of SDS.

3. CONTENT-DEPENDENT SECURITY

Despite several security aspects of SDS were analyzed in [7], we will focus on the resistance of an SDS to forgery attack in this section.

3.1. Resistance to Disjoint-Block Counterfeit Attacks

The counterfeit attacks [2, 4] are based on block search and match. However, when the same operation is conducted to defeat the wavelet-based authentication scheme [7], it is quite infeasible to expect the desired performance because the multiscale structure of wavelets contradicts the assumptions of [2, 4]. In the wavelet domain, each parent-child pair maps to a set of spatial pixels, which is of *non-fixed* size and possesses certain contextual dependencies. Under these circumstances, there is no clue (how to determine the size of a block?) that can be utilized to make the collage attack success. This explains the advantage of adopting the multiscale structure of wavelets in designing an image's signature.



(a) (b)

Fig. 1. Illustration of some parent-child pairs of an SDS in the wavelet domain. Two-level wavelet decomposition is performed. (a) Any two nodes connected by a line is a pair with their magnitude difference larger than a threshold ρ (Table 1). (b) Significant parent-child pairs are mapped back into the pixel domain as isolated regions (illustrated in normal gray-scale). Apparently, the structural contents (e.g., textures and edges) are preserved.

Some results are shown in Fig. 2 and Table 2 to demonstrate the unique anti-forgery of structural information. Figs. 2(a) and (b), respectively, show the counterfeit Lenna images generated from a database based on 32×32 and 8×8 block search and match. Our database is composed of 1000 images, excluding the Lenna. By inspecting Fig. 2(b) and original Lenna, they look perceptually similar but blocky effects remain. By comparing the structural digital signatures extracted from counterfeit images with that extracted from the original Lenna, the bit error rates (BERs) are listed in Table 2. If the counterfeit image is created in an 8×8 basis and its SDS is extracted based on $\rho = 256$, the BER (as high as 0.276) is the smallest in Table 2. This is because only a few lower frequency features, which is effective in hiding differences, are captured. However, the resultant BER is still sufficient to reveal that most regions have been malicious tampered with. In sum, these results verify that an SDS indeed resists against block-based counterfeit attacks.

Table 2. **BERs measured between the original and counterfeit images. One bit error means a different parent-child relationship.**

forged images	32×32(2(a))	16×16	8×8(2(b))
$\rho = 64$	0.771	0.591	0.337
$\rho = 256$	0.785	0.548	0.276

Table 1. Construction of Structural Digital Signature (SDS)

1.	Compute the DWT of an image. In our implementation, the size of the lowest frequency band is fixed to be 16×16.
2.	Select those parent-child pairs with their magnitude difference larger than a pre-determined threshold ρ. We consider this kind of pairs *significant*. In fact, ρ is determined from the desired false positive and false negative probabilities.
3.	For each selected pair, $< p, c >$, it is classified as one of four types defined as follows. Type I: $p > 0$ and $\|p\| > \|c\|$; Type II: $p < 0$ and $\|p\| > \|c\|$; Type III: $c > 0$ and $\|p\| < \|c\|$; Type IV: $c < 0$ and $\|p\| < \|c\|$.
4.	Initially, SDS$[i, j]$=V for $\forall \ \ i, j$. The SDS array is recorded as SDS$[i, j]$=I or II or III or IV according to step 3, where $[i, j]$ is a child's coordinate of a significant pair in the wavelet domain.

(a) (b)

Fig. 2. Block-based counterfeit attack: (a) counterfeit Lenna (22.33dB) based on 32×32 block searching; (b) counterfeit Lenna (28.51dB) based on 8×8 block searching.

3.2. Resistance to Forgery under Structural Information Leak

As we have described in Table 1, an SDS is constructed from the magnitude relationships of significant parent-child pairs in the wavelet domain. In other words, an SDS has been created by taking larger wavelet coefficients into consideration. As described in [9], larger wavelet coefficients can be efficiently used to approximately reconstruct an original signal. Now, we will prove that the proposed SDS can tolerate the Radhakrisnan and Memon's attack [10].

Owing to an SDS is constructed from the parent-child pairs with significant magnitude differences, locations of these significant pairs in an image are not a secret at all so that attackers can know this prior knowledge. Let Y be a target image from which attackers try to create an SDS similar to that of an original image X by using this publicly known prior knowledge. The only way that attackers can do is to modify the magnitude relationships of parent-child pairs (which are significant in X) to be significant. In addition, the modified pairs still have to be selected in the verification stage without affecting the authentication capability. If the pair-wise wavelet coefficients in Y have been modified but cannot be selected later for verification, then this kind of modifications is regarded as useless. Consequently, it is extremely difficult to enforce an undesired pairwise relationship while maintaining transparency. If an image to be authenticated is of poor quality, then it can be directly rejected before it is fed into an authentication system.

Let Y be modified as Y^*. There may exist some ways for attackers to preserve the same SDS between X and Y^*. A simple way is to first sort the wavelet coefficients of images X and Y, respectively. Then, the sorted result of Y is re-ordered according to the sorted result of X and is reconstructed by inverse wavelet transform as Y^*. It is said that this attack is successful if Y^* is sill perceptually similar to Y. With this attack, we have the following corollary.

Corollary 1: Let X and Y be two different images and Y^* be obtained by means of forgery of Y's structural information. The perceptual similarity among the three images will satisfy $PSNR(Y, Y^*) << PSNR(X, Y^*)$.

Proof: An image Y is modified as Y^* such that both X and Y^* have the same parent-child relationships among the selected significant pairs. In addition, based on the principle of [7] the selected parent-child pairs are more significant than those that are not selected. As a result, an image X can be specified as a significant term $S(X)$ and a residual term $R(X)$, i.e., $X = S(X) + R(X)$. When a forgery attack (see the description given above) is applied on Y to obtain Y^*, we have $Y^* = S(X^*) + R(Y)$, where $S(X^*)$ and $S(X)$ denote the signals reconstructed from the same significant parent-child pairs (they may have slight difference in magnitudes). Furthermore, according to the signal reconstruction mechanism [9], the reconstructed signal Y^* is perceptually similar to X because both of them have similar semantic features, i.e., $S(X^*) \approx S(X)$. In practice, attackers may wish to discard all un-selected pairs or make them small such that the magnitude difference of a selected "significant" pair could be not large. Under this circumstance, the reconstructed signal, $S(X^*)$, will still be a smooth version of its original one, X. Besides, the high-frequency components can also be retained for image reconstruction. This will act like noise addition and will not affect the perceptual similarity between X and Y^*. On the other hand, the perceptual similarity between Y and Y^* will be low because their significant part, $S(Y)$ and $S(X^*)$, are greatly different [9]. Based on the above deductions, we have verified the result: $PSNR(Y, Y^*) << PSNR(X, Y^*)$.

This corollary states that it is impossible to simply simulate an image Y^*, which is modified from Y to have the same multiscale structure of an image X and is still perceptually similar to Y. Some results are shown in Fig. 3 to further verify the above corollary. It can be observed from the forged images (Figs. 3(c)$\sim$(f)) that the perceived structural information is changed remarkably. Undoubtedly, significant parent-child pairs (structural information) dominate the image's content.

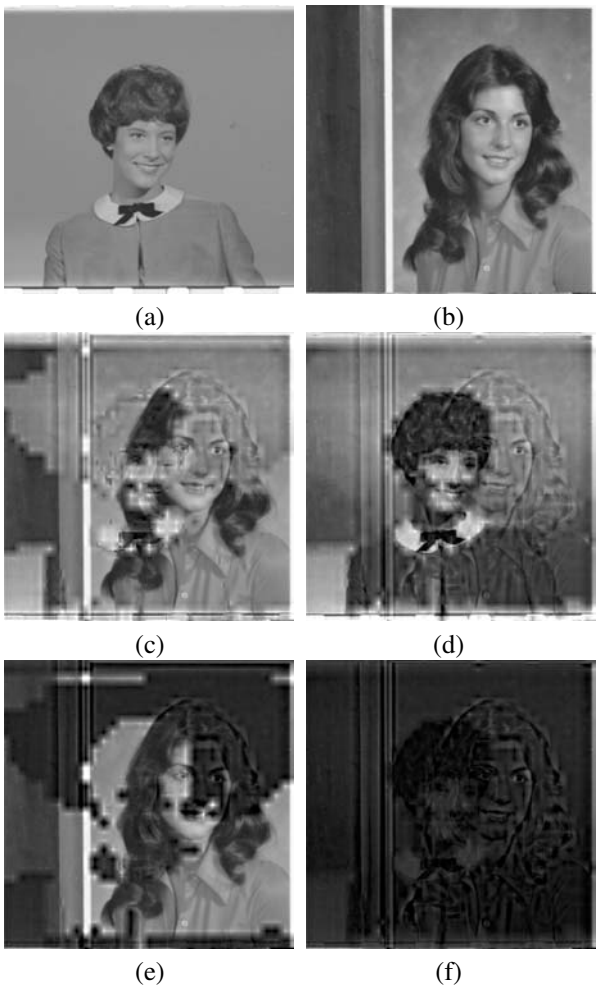


Fig. 3. SDS copy: a target image is modified to have the same significant pairs that an original image has. (a) an original image; (b) a target image; (c) and (d) are generated by sorting the wavelet coefficients and then placing the first 500 and 2000 larger magnitudes of (b) to corresponding locations determined from (a), respectively; (e) and (f) are modified from (c) and (d), respectively, with magnitudes of selected wavelet coefficients scaled to a quarter size. With SDS copy, the modified target images ((c)$\sim$(f)) are gradually degraded and similar to the original image (a).

4. CONCLUSION

This paper explores to extract or embed structural information for image authentication. The security of the content-dependent structural digital signature has been particularly emphasized. Our analyses confirm that under counterfeit attacks (i) structural features demonstrate strong robustness and security; (ii) block-based features are easy to be forged. This is because wavelet-based structural information inherits its multiresolution characteristic to break the independency inherent in conventional disjoint block-based (single resolution) transforms.

5. REFERENCES

[1] R. W. Buccigrossi and E. P. Simoncelli, "Image Compression via Joint Statistical Characterization in the Wavelet Domain," *IEEE Trans. on Image Processing*, Vol. 8, No. 12, pp. 1688-1701, 1999.

[2] J. Fridrich, M. Goljan, and N. Memon, "Further Attacks on Yeung-Mintzer Fragile Watermarking Scheme," *Proc. SPIE: Security and Watermarking of Multimedia Contents*, 2000.

[3] M. Holliman, N. Memon, and M. Yeung, "On the Need for Image Dependent Keys in Watermarking," *Proc. of the Second Workshop on Multimedia*, NJIT, 1999.

[4] M. Holliman and N. Memon, "Counterfeiting Attacks on Oblivious Block-wise Independent Invisible Watermarking Schemes," *IEEE Trans. on Image Processing*, Vol. 9, No. 3, pp. 432-441, 2000.

[5] C.-Y. Lin and S.-F. Chang, "A Robust Image Authentication Method Distinguishing JPEG Compression from Malicious Manipulation", *IEEE Trans. on Circuits and Systems for Video Tech.*, Vol. 11, pp. 153-168, 2001.

[6] C. S. Lu, "Wireless Multimedia Error Resilience via a Data Hiding Technique," *Proc. IEEE Int. Workshop on Multimedia Signal Processing*, USA, 2002.

[7] C. S. Lu and H. Y. Mark Liao, "Structural Digital Signature for Image Authentication: An Incidental Distortion Resistant Scheme," *IEEE Trans. on Multimedia*, Vol. 5, No. 2, pp. 161-173, 2003.

[8] C. S. Lu and C. Y. Hsu, "Content-dependent Anti-Disclosure Image Watermark," *Proc. Int. Workshop on Digital Watermarking*, LNCS 2939, Korea, 2003.

[9] S. Mallat and S. Zhong, "Characterization of Signals from Multiscale Edges," *IEEE Trans. on Pattern Anal. and Machine Intell.*, Vol. 14, No. 7, pp. 710-732, 1992.

[10] R. Radhakrisnan and N. Memon, "On the Security of the SARI Image Authentication System," *Proc. IEEE Int. Conf. on Image Proc.*, Vol. III, pp. 971-974, 2001.

[11] P. W. Wong, "A Public Key Watermark for Image Verification and Authentication," *Proc. IEEE ICIP*, pp. 425-429, 1998.