Tampering attacks on binary phase only filter–based watermarking schemes for image authentication

Gang Cao
Yao Zhao
Rongrong Ni
Tampering attacks on binary phase only filter–based watermarking schemes for image authentication

Gang Cao
Yao Zhao
Rongrong Ni
Beijing Jiaotong University
Institute of Information Science
No. 3 of Shangyuan Residence Haidian District
Beijing, Beijing 100044, China
and
Beijing Key Laboratory of Advanced Information Science and Network Technology
Beijing 100044, China
E-mail: 06112056@bjtu.edu.cn

Abstract. Recently, a category of watermarking techniques based on binary phase-only filter (BPOF) has been proposed for image authentication. In such techniques, the authentication is implemented by evaluating the correlation between Fourier phase information and the hidden watermark bitplane. In this paper, we reveal the security flaws of BPOF-based watermarking algorithms and propose sophisticated tampering attacks against them. We show how the attacker can easily tamper with a watermarked image without being detected. Experimental results demonstrate that our attacks are successful in tampering watermarked images. The watermarking schemes are proven to be fundamentally flawed.

Subject terms: digital watermarking; image authentication; watermarking security; attack; image tampering; phase spectrum; magnitude spectrum.

Paper 100620RR received Jul. 30, 2010; revised manuscript received Mar. 21, 2011; accepted for publication Mar. 28, 2011; published online May 20, 2011.

1 Introduction

In today’s digital age, digital image manipulation becomes much easier and prevalent. Seeing from digital images becomes no longer believable. It becomes significant to verify the content integrity and authenticity of digital images. The watermarking-based image authentication technique is developed as one solution for such a requirement. Many published authentication watermarking schemes are in Refs. 1–12.

The basic task of a watermark-based image authentication system is to identify whether the image data has been tampered. Unacceptable distortion on the data should be alerted while legitimate manipulation should be permitted. An authentication scheme usually consists of protection and verification processes. For protection, the watermark, which can be considered as a compact representation of image content, is generated and hidden into the image data itself. For verification, the hidden watermark is extracted and compared with the regenerated one. Inferior matching between them implies the occurrence of malicious tamper.

Recently, a self-authentication watermarking scheme based on Fourier phase signature has been proposed in Ref. 1, then improved in Ref. 2, and widely applied in Refs. 3–5. The BPOF watermark is generated from an image’s Fourier phase spectrum and then hidden back into the magnitude spectrum. A watermark detector evaluates the correlation between phase information and the extracted BPOF watermark to identify whether the received watermarked image has been tampered or not. The BPOF-based watermarking scheme primarily proposed in Ref. 1 is referred to as the basic BPOF scheme throughout this paper. To improve the authentication’s reliability, Sang et al. proposed to use only the low-frequency range of the Fourier spectrum to generate and embed a watermark, which is referred to as the improved BPOF scheme.

For an authentication watermarking scheme, security refers the ability to resist intentional tampering by opponents in the channel and is measured by the opponents’ inability to launch a successful attack. A watermarking algorithm is secure if it can alarm malicious tampering and authenticate unaltered images. Watermarking security brings new challenges in the image authentication field. Security of the watermarking algorithms resisting different types of well-designed attacks is extensively explored. In this paper, we focus on the security analysis of BPOF-based watermarking schemes. We show how they can be defeated by our proposed attacks, which manage to exploit the redundancy in such tampering space: Fourier phase spectrum, Fourier magnitude spectrum, and spatial domain, respectively. The attacks succeed in destroying or forging the image’s content while deceiving the watermark detectors. The destructive attacks including phase and magnitude tampering may be used by malicious users to destroy the image content, which impairs the users’ benefit. The destructively attacked images can fool the authentication system, especially in the case of large batch automatic processing and without subjective human inspection. Forging an attack in the spatial domain can be applied to create a new forgery image with different semantic meaning, which not only deceives the authentication system but also fools the human eyes.

The rest of this paper is organized as follows. In Sec. 2, we briefly review BPOF-based watermarking schemes. Security flaw of such schemes is deeply explored and three different types of attacks are proposed in Sec. 3, followed by the report of experimental results in Sec. 4. The conclusions are drawn in Sec. 5.

2 BPOF-Based Watermarking Schemes

In this section, we first review the basic BPOF-based watermarking scheme including two stages: watermark embedding and watermark detection, followed by the introduction of the improved BPOF scheme.

2.1 Watermark Embedding

In the basic BPOF scheme, the watermark embedding process is defined as follows.
To authenticate the received image, the original image is transformed into discrete Fourier transform (DFT) domain. That is, 
\[ I(u, v) = X(u, v)e^{j\phi(u, v)}, \]

where \( u, v \) represents the spatial frequency coordinates. \( X(u, v) \) is the magnitude of frequency coefficient \( I(u, v) \). Phase of the Fourier spectrum is denoted by \( \phi(u, v) \), in the range \((-\pi, \pi)\).

Magnitude is rounded to the nearest integer value and represented by bitplanes:
\[
R(u, v) = \text{round}(X(u, v)) = \{R_q, R_{q-1}, \ldots, R_k, \ldots, R_2, R_1\}. \tag{1}
\]

In this formula, \( R_k \) denotes the \( k \)th bitplane of rounded magnitude. \( q \) is the sequence number of the highest-level nonzero bitplane.

BPOF is calculated from the phase spectrum by first imposing binary conversion:
\[
b(u, v) = \begin{cases} +1 & \text{if } \cos \phi(u, v) \geq 0 \\ -1 & \text{otherwise} \end{cases}. \tag{2}
\]

Next the bipolar binary pattern is mapped to a unipolar pattern \{1, 0\}, which is the BPOF of the image and denoted as \( B(u, v) \). To prevent spoiling and provide security, the original BPOF is encrypted by a non-avalanche encryption function \( E[\cdot; K] \) with the key \( K \). The encrypted BPOF is \( E_b(u, v) = E[0, v; K] \).

BPOF is computed by
\[
E_b(u, v) \text{ is embedded by substituting one bitplane of the rounded magnitude } R(u, v) \text{ to be kept invariant, while content of the watermarked images is distorted as much as possible. The framework of the proposed attack system is shown in Fig. 1. According to the difference of tampering domains, three attack methods are designed based on phase tampering, magnitude tampering, and spatial-domain tampering, respectively. The attack in each method is implemented by}
\]

Correlation between \( b_T(u, v) \) and \( \Phi \) is computed by performing IDFT as follows:
\[
\text{Corr}(u, v) = \text{IDFT}(b_T(u, v) \cdot \Phi(u, v)). \tag{5}
\]

The correlation is quantified by two metrics: peak-to-average-correlation energy (PACE) and peak-to-secondary-peak ratio (PSR), which are defined as
\[
\text{PACE} = 20 \log_{10} \left( \frac{P_{\max}}{\mu} \right) \tag{6}
\]
\[
\text{PSR} = 20 \log_{10} \left( \frac{P_{\max}}{P_{\text{second}}} \right). \tag{7}
\]

Here, \( P_{\max} \) and \( P_{\text{second}} \) indicate the highest and the second highest peak of the correlation plane \( \text{Corr}(u, v) \) respectively. \( \mu \) means the average value of the correlation plane. Based on such metrics, an authentication decision can be made by the thresholding method. Images with higher metric would be judged as authentic ones, while lower metric values signify forgeries or tampered ones.

### 2.3 Improved BPOF-Based Watermarking

An improved BPOF-based watermarking method was proposed in the recent literature. In such an improved scheme, only the Fourier spectrum in a low-frequency range was employed to embed and detect watermark. Illustrative experimental results gained on the standard test images verified that the improved scheme could achieve excellent semifragile performance if low frequencies between 25×25 and 49×49 and high-level magnitude spectrum bitplane were applied for an embedding watermark.

Besides, another metric called identical ratio (IR) has been proposed to detect watermark and measure authenticity in Ref. 2. IR is defined as the identical ratio of bits between the extracted watermark bitplane and the recomputed BPOF bitplane. Comparing the correlation-based metric PACE and PSR, IR is computationally efficient and has better watermark-detection performance.

### 3 Proposed Attacks

In the BPOF-based watermarking schemes, the secure encryption mechanism is introduced during watermark generation. In fact, it has been successful to prevent unauthorized authentication. Because the authentication can be performed only when the extracted BPOF watermark bitplane is decrypted by the authorized users who have the key. However, the inappropriate watermark generation and embedding methods leave behind potential security loopholes for sophisticated attackers. Here, we propose three types of well-designed tampering attacks. We will show how the attacked images escape the detection of BPOF watermarking system.

The objective of the proposed attacks is that both recomputed BPOF bitplane and the embedded watermark bitplane are managed to be kept invariant, while content of the watermarked images is distorted as much as possible. The framework of the proposed attack system is shown in Fig. 1. According to the difference of tampering domains, three attack methods are designed based on phase tampering, magnitude tampering, and spatial-domain tampering, respectively. The attack in each method is implemented by
two stages: tampering and correction. The three methods are integrated here for illustration and are separately applied.

The proposed attacks are targeted for tampering the BPOF-based watermarked images without being detected. So it is required to preconceive that BPOF watermarking has been applied in the targeted images. For the BPOF watermarking methods, the location of the watermark can be recomputed exactly. Furthermore, we consider the more complicated situation: in case an unfixed embedding position is used to protect their watermark. To deal with the challenge of an unknown watermark bitplane, we propose the corresponding blind attack method in Sec. 3.4.

3.1 Attack Based on Phase Tampering

Although Fourier phase of a digital image is highly content-dependent and sensitive to image alteration, the authentication schemes still leave redundant modification space for attackers. Recall that the BPOF watermark $B(u, v)$ is generated from the binary phase filter given in Eq. (2). We can see that each element of the watermark bitplane is determined by phase polarity. If only the phase falls into the range $[-\pi/2, \pi/2]$, the generated watermark bit would always ‘1.’ Contrarily, out of this range means that the watermark bit would be ‘0.’ An intuitive inference is that the phase can be modified as long as the falling range is kept invariant.

Solely based on such considerations, we present a phase tampering method to attack the basic BPOF watermarking scheme as follows:

(C1) The received image $i_T(x, y)$ is transformed to $I_T(u, v) = X_T(u, v)e^{j\phi_T(u, v)}$ by DFT.

(C2) The phase is modified according to the following regulation:

$$
\phi_T'(u, v) = \begin{cases} 
\phi_T(u, v) + \alpha, & \text{if } \phi_T(u, v) \in \{[-\pi/2, 0] \cup (\pi/2, \pi)\} \\
\phi_T(u, v) - \alpha, & \text{if } \phi_T(u, v) \in \{(0, \pi/2] \cup (-\pi/2, -\pi/2)\} \\
\phi_T(u, v) - \alpha/2, & \text{if } \phi_T(u, v) = \pi 
\end{cases}
$$

(8)

where $\alpha \in (0, \pi/2]$ is the shift angle for each quadrant. Tampering the phase in the fourth quadrant is illustrated in Fig. 2.

(C3) The initially tampered image $i_{A0}(x, y)$ is acquired by first doing IDFT on the altered phase $\phi_T'(u, v)$ and untouched magnitude $X_T(u, v)$, followed by being truncated and rounded into a normal image data range, i.e., $[0, 255]$ for 8-bit grayscale image.

(C4) DFT is done on $i_{A0}(x, y)$ and we have $I_{A0}(u, v) = X_{A0}(u, v)e^{j\phi_{A0}(u, v)}$.

(C5) Magnitude $X_T(u, v)$ and $X_{A0}(u, v)$ are represented by bitplanes, respectively,

$$
R_{T}(u, v) = round(X_T(u, v)) = \left\{ R_{T_0}^T, R_{T_{q1-1}}^T, \ldots, R_{T_{q2-1}}^T, R_{T_0}^T \right\}.
$$

(9)

$$
R_{A0}(u, v) = round(X_{A0}(u, v)) = \left\{ R_{A0}^{q1}, R_{A0}^{q1-1}, \ldots, R_{A0}^{q2}, R_{A0}^{q2} \right\}.
$$

(10)

Here, $q_1$ and $q_2$ are the sequence number of the highest nonzero bitplane in round($X_T(u, v)$) and round($X_{A0}(u, v)$), respectively. Then, the $h$th bitplane of $R_{A0}(u, v)$ is replaced by the $h$th bitplane of $R_T(u, v)$,

$$
R_{A0}(u, v) = \left\{ R_{A0}^{q1}, R_{A0}^{q1-1}, \ldots, R_{A0}^{q2}, R_{A0}^{q2} \right\}.
$$

(11)

where $h = \left[ \frac{q_1+1}{2} \right]$.

![Fig. 1 Framework of the proposed attack system.](image-url)
In the tampering stage (C1)–(C3), phase is shifted toward the neighboring quadrant in the same polarity space to keep recomputed BPOF invariant after attack. Shift angle $\alpha$ can be chosen to get the desired effect for destroying image content. Generally, the resulting image degrades more seriously if a larger angle is shifted. However, the gray level of more inverse-transformed pixels goes beyond the normal range when shifting a larger angle. Once truncated, such pixel values will be seriously distorted. Then, the BPOF bitplane recomputed from the initially tampered image $i_{A0}(x, y)$ presents a much greater difference to that of the original image. That is not beneficial to keep the BPOF bitplane invariant and matched with the hidden watermark bitplane. So there exists a tradeoff between the shift angle and the attack’s reliability.

On the other hand, we strive to keep the embedded watermark bitplane as invariant as possible. To this end, a correction strategy is exploited as the steps (C4)–(C7). In fact, magnitude spectrum of $i_{A0}(x, y)$ becomes completely different to that of the untampered image $i_T(x, y)$, especially when a larger shift angle is applied. The corresponding watermark

(C6) The corrected tampered image $i'_{A0}(x, y)$ is gained by doing IDFT on $R_{A0}(u, v)$ and $\phi_{A0}(u, v)$.

(C7) Let $i_{A0}(x, y) = i'_{A0}(x, y)$, repeat steps (C4)–(C7) for $N$ times until generating the attacked image $i_A(x, y) = i'_{A0}(x, y)$.

Fig. 2 Tampering the phase in the fourth quadrant.

Fig. 3 Metric of phase tampered images in the basic BPOF scheme, including (a) PACE metric, (b) PSR metric, (c) IR metric, and (d) PSNR (dB) metric. Numbers 1 to 23 refer to images in the first set ($512 \times 512$) and numbers 24 to 36 refer to images in the second set ($256 \times 256$), which is the same in the following metric figures.
biplanes also mismatch. So we need to correct the distorted watermark biplane of $i_{A0}(x, y)$. Fortunately, the embedding position of the original watermark biplane $h$ can be easily recomputed according to Eq. (3). Because the embedding operation does not change the number of biplanes from which the image can be decomposed, we can accurately relocate and extract the watermarked biplane before tampering. Motivation for repeating the re-embedding process is to enhance the coherence between the watermark biplane extracted from $i_T(x, y)$ and that from $i_A(x, y)$, while not decreasing the consistency of the BPOF biplanes. Hence, the attacked image $i_A(x, y)$ can deceive the BPOF watermark detector because the matching between the recomputed BPOF biplane and the extracted watermark biplane is still maintained and without any obvious destruction.

The same attack method can be applied to the improved BPOF watermarking scheme. In the tampering stage, we can use a larger shift angle even $\alpha = \pi/2$. That is because phase correction within low frequencies is performed in the subsequent correction stage, which can ensure the invariance of the recomputed BPOF biplane more strictly. The low-frequency-range phase of the initially tampered image $i_{A0}(x, y)$ is replaced by that of the original image during each iterative correction. Simultaneously, we repeatedly reembed the watermark biplane merely in the corresponding low-frequency range.

### 3.2 Attack Based on Magnitude Spectrum Tampering

Besides the phase, magnitude spectrum is also an important component for representing an image’s content. It can reflect the spatial structure and texture characteristics of an image. Hence, we can destroy image content by modifying its Fourier magnitude spectrum. The second attack is realized by two vital stages: magnitude tampering and tampered image correction.

In the magnitude tampering stage, we modify high-level magnitude bitplane(s) of $i_T(x, y)$ to be all-zero bitplane(s) excluding the highest-level nonzero bitplane. Specifically, we alter $R_T^r(u, v)$ to be

$$R_T^r(u, v) = \{R_{q1}^r, R_{q1-1}^r, \ldots, R_{q}^r, \ldots, R_{q}^r, R_{q0}^r\}. \quad (12)$$

Without loss of generality, we let

$$R_{q0}^r = \left\{ \begin{array}{ll} \theta & \text{if } k \in [1, q_1 - N_0] \\ R_{q1}^r & \text{otherwise} \end{array} \right. \quad (13)$$

Here, $N_0 \in \{1, 2, 3, \ldots, q_1 - 1\}$ is the number of unchanged biplanes. The all-zero bitplane is denoted by $\theta$. The initially tampered image $i_{A0}(x, y)$ is yielded by doing IDFT on the altered magnitude $R_T^r(u, v)$ and the untouched phase $\phi_T(u, v)$.

Reasons for such modification are explained here. The aim for excluding the highest bitplane is to ensure the number of decomposed magnitude bitplanes invariant after tampering. That respects the principle for selecting an optimal embedding position in primal watermarking schemes. The choice of setting zero is for the sake of reducing truncation error that occurs in IDFT. Meanwhile, more zeros present in high biplanes can blacken the image and demolish content details. Setting which and how many biplanes to be zero can be concretely designated by attackers according to their expectation on the extent of distortion.

In the subsequent stage of correction, watermark bitplane is re-embedded as the step (C5) in phase tampering attack, which aims to keep itself invariant. Different from the previous case, phase of the original image $\phi_T(u, v)$ should be inherited and applied in inverse DFT during each iterative correction. Corresponding to step (C6) in the phase tampering method, the corrected tampered image $i_T^r(x, y)$ is generated by doing IDFT on $R_A(u, v)$ and $\phi_T(u, v)$, instead of $\phi_A(u, v)$. Owning to the maintenance of original phase information, the recomputed BPOF bitplane from attacked images resembles that from original images. So a higher authentication metric can be gained. The BPOF watermark detector would misjudge such attacked images as authentic ones, indeed they have been changed beyond recognition.

For the improved BPOF-based watermarking scheme, the same modification can be done to magnitude bitplanes. Then, both phase updating and watermark biplane re-embedding are performed in the low-frequency range, the same as in the previous case. Because pixels’ variability induced by such
correction is smaller than that of the basic BPOF scheme, a more reliable attack can be achieved.

3.3 Attack Based on Spatial-Domain Tampering

In fact, the image tampering can be classified into two kinds: destructive and forging tampering. The modification of Fourier phase and magnitude spectrum belongs to the destructive attack that destroys the visual content of an image. Although the destructively attacked images can be easily noticed by human eyes, they can still pass the automatic image authentication system. This works especially in the case of large batch automatic authentication, where it is not practical to observe the images one by one by human eyes. Forging tampering is to change some specific details of an image and create a forgery image with different semantic meaning. In this section, the attack based on spatial-domain tampering only belongs to forging tampering. Such a tampering attack not only deceives the authentication system but also fools the human eyes, so it would be more advantageous for the adversary.

In the first stage, attackers are allowed to edit the content of an image through the pixel-domain tampering. Objects in the image can be removed, added, and retouched arbitrarily. In the second stage, the same correction as the steps (C4)–(C7) in the phase tampering method is performed. It is necessary to point out that phase information of the original watermarked image should not be inherited. That attributes to the obvious phase difference between the original watermarked image and its spatially tampered version, especially when a larger region has been altered. If the original phase information is forcibly inherited, the resulting attacked image will be distorted by phase disturbance and lacks realistic visual quality.

For attacking the improved BPOF scheme, we can first edit the image appearance. Subsequently, the correction without a phase update is employed. Only the watermark bitplane is iteratively re-embedded in the low-frequency range.

3.4 Challenge of Unknown Watermark Bitplane

In the above three attack methods, the watermark bitplane can be exactly relocated according to Eq. (3). However, the authentication schemes may be improved to embed the BPOF watermark into an unknown magnitude bitplane, which is not in terms of the determinate rule. The specific position of watermark-embedding may be taken as a key communicated between legal users. In such a scenario, attackers fail to easily relocate the watermark bitplane.

Even so, we can confirm that the BPOF watermark bitplane must have been hidden in a certain middle-level (i.e., 12th–14th) bitplane. This fact is determined by the semifragility requirement of BPOF-based watermarking schemes. In this case, we can amend the previous attack methods by borrowing ideas from a blind attack. The countermeasure is to simultaneously update the multiple middle-level magnitude bitplanes during each iterative correction. Specifically, several middle-level bitplanes of the initially tampered image \( R_\alpha(u, v) \) are replaced by those of the original image \( R_T(u, v) \). As for the improved BPOF scheme, only the low-frequency range of multiple middle-level bitplanes is required to be updated.

4 Experimental Results

In this section, simulation results of the proposed tampering attacks are reported. To guarantee impartiality, evaluation of the proposed attack is performed on the test data that have been used in the experiments of BPOF-based watermarking algorithms. The test data includes two sets of natural images obtained from the USC SIPI database. The first set contains 13 images with sizes of 256\( \times \)256, while the second set has 23 different 512\( \times \)512 images. The dataset consists of scenery, aerial, texture, and motion frame images. Note that some of the images are color images and we convert them into YCbCr format first, then truncate and round the \( Y \) channel to be corresponding grayscale versions. The results furnished in this section are obtained by testing on the grayscale images.

In the following experiments and if without specification, the related parameters are set as follows. To simulate the BPOF watermarking algorithms, the BPOF watermark bitplane is embedded into the \( h \)th magnitude bitplane strictly
4.1 Evaluating the Phase Tampering Attack

To evaluate the phase tampering attack, we compare the authentication metric and image quality metric of watermarked images before and after an attack. As shown in Figs. 3(a)–3(c), three kinds of authentication metric on each test image have been given, respectively, which have consistent measure effect. In each subfigure, metric computed from original images, watermarked images, and tampered watermarked images are plotted, respectively. We can see that the authentication metric of the attacked images are much closer to that of the watermarked images, even for a larger phase shift angle $\alpha = 3\pi/8$. According to Ref. 2, if the experimental threshold of IR is set as 0.65, PACE/PSR are set as 40/20, most of the tampered images are misjudged as authentic by the BPOF watermark detector. While the attacked images gain high authentication metric, image quality has degraded badly. That can be verified by the lower PSNR values reported in Fig. 3(d) and example images shown in Fig. 4.

However, there also exists unsuccessful attack on some images in the test set [see Fig. 3(c)]. The phase tampering attack ($\alpha = 3\pi/8$) on the images with Nos. 30 and 31 are not successful because the IR values of the attacked images are much lower than the threshold of 0.65. The unsuccessfulness attributes to an inherent low metric of the watermarked images themselves, which is determined by the limitation of the watermarking algorithm itself. A larger attack strength might also result in more risk for unsuccessfulness.

Evaluation results of the improved BPOF scheme are shown in Figs. 5 and 6. Compared with attacks on the basic BPOF scheme, higher IR metric and comparative destruction are achieved because the largest shift angle ($\alpha = \pi/2$) and low-frequency correction are allowed to be performed. As a result, we can conclude that the attack based on phase tampering is successful. It can deceive both the basic and improved BPOF authentication systems.

4.2 Evaluating the Magnitude Tampering Attack

Results for assessing the magnitude tampering attack are shown in Figs. 7 and 8. Here, the number of untampered magnitude bitplanes $N_0$ is set as 1, 6, and 11, which represent three different levels of attack strength: intense, moderate, and weak, respectively. This can be observed from the examples in Fig. 8 and the PSNR value in Fig. 7(a).
Fig. 9 Forgery images in the basic BPOF scheme. The three rows show watermarked images, forgery images, and corrected forgery images, respectively. PSNR (dB) between the forgery image and its corresponding corrected version is shown in the last row.

\[ N_0 = 1 \], the highest bitplane is untouched while all other bitplanes are set as zero. The attacked images are black and do not carry any scene information.

In Fig. 7, we can see that IR metric of most attacked images are higher than that of unaltered watermarked images, far above the threshold of 0.65. Moreover, variability of the IR metric is quite small via different \( N_0 \), especially for 512×512 images. The attack on the improved BPOF scheme is also testified to be efficient by experimental results. By comparing the performance between the phase and magnitude tampering attacks, we can find that the later one behaves more excellently. When the anticipative tampering effect is acquired, a higher authentication metric can also be ensured. Moreover, the attacks on the improved BPOF scheme behave better than those on the basic BPOF scheme.

4.3 Evaluating the Spatial-Domain Tampering Attack

In this subsection, we test the effectiveness of the spatial-domain attack via practical forgery samples. First, the original images of “Lena” (512×512), “Peppers” (512×512), and “House” (256×512) are watermarked by the basic BPOF authentication method. Then the watermarked images are manipulated in Photoshop CS4 software and counterfeited to be forgery images, which are shown in Fig. 9. For the watermarked Lena image, we copy the flower and chaplet from other internet images and paste onto Lena’s head and hair, respectively. Here, internet images are used to ensure the universality of the forgery creating process. For the watermarked Peppers image, five copies of one capsicum region is pasted within the same image. For the watermarked House image, the chimney is removed and the window is copied and pasted twice. Graceful realism can be perceived from the tampered watermarked images.

As shown in the middle row of Fig. 9, the tampered watermarked images are the direct output from Photoshop and without post correction. Their IR metric are lower than the threshold of 0.65, so they would be identified as unauthentic images by the BPOF watermark detector. However, once being processed by our proposed correction strategy in the spatial-domain tampering technique, such forgery images’ IR values become larger and exceed the threshold. Consequently, the corrected forgeries can pass the authentication and fool the watermarking system. Successfulness of the attack is at the cost of venial degradation of the image quality, which can be perceived by the visual difference and PSNR values between the uncorrected and corrected forgery images.

To simulate the attack on improved BPOF scheme, as that in Ref. 2, one 128×128 random noise block is overlapped onto watermarked images shown in Fig. 10. The random noise obeys a uniform distribution over the interval [0, 255]. Satisfying results are gained by the correction-based tampering method. Better quality of tampered images can still be obtained when a higher IR metric is attained, especially for texture images such as “Baboon.” In fact, the watermarking method can only resist 8×8 pixels tampering.2 We have to admit that the proposed spatial-domain attack method can only enlarge the freedom of tampering operations to a certain extent. If too large region is tampered on a nontextured image, our method might be invalidated due to the intense variability of the image phase.
4.4 Evaluation in the Case of Unknown Watermark Bitplane

In this section, we test our proposed attack methods in the case of an unfixed embedding position. In this situation, attackers do not know which magnitude bitplane has hidden the watermark. So they can simultaneously inherit multiple middle-level bitplanes. In our experiments, we blindly re-embed the 11th–15th magnitude bitplanes in correction stage. Without loss of generality, the magnitude tampering attack is chosen for testing. \( N_0 \) is set as six for simulating a moderate attack strength. The watermark is supposed to be embedded into the 12th, 13th, or 14th magnitude bitplane. Results for attacking the basic BPOF scheme are shown in Fig. 11. It can be seen that the authentication metric of attacked images closely follows that of watermarked images. It can also be found that if only the watermarked image is able to pass authentication (i.e., IR > 0.65), the corresponding attacked image would be wrongly detected as authentic. At the same time, the desired distortion effect has been achieved, which can be validated by the low PSNR values. We can note that the fluctuated and lower metric of the watermarked images is due to the limitation of the watermarking algorithm itself.

Corresponding results for the improved BPOF scheme are reported in Fig. 12. Compared with the basic BPOF scheme, the results are less sensitive to image content and size. For the case of \( b = 12 \), IR metric values of 512×512 test images are much lower, around 0.5 to 0.6. That is because the watermarking algorithm fails to support embedding watermark in \( b \)th level bitplane, which is also determined by the algorithm’s limitation. Even so, the attack scheme can keep the IR metric nearly invariant.

5 Conclusion

In this paper, we investigate the security of the BPOF-based authentication watermarking schemes.\(^1\,^2\) We have demonstrated how to tamper with both the Fourier spectrum and spatial-domain pixels of a watermarked image. While content if the tampered images is destroyed or forged, a high authentication metric can still be obtained. So, the watermark detectors can be successfully deceived and fail to identify the attacked images. Experimental results have verified the successfulness of our proposed attacks.

We conclude that current BPOF-based watermarking schemes are not secure for its designed application of digital image content authentication. We suggest that authors reconsider the security issues and redesign the watermarking schemes before they can be used for practical applications. A fatal deficiency of those watermarking methods lies in the transparent generation of the BPOF in step (A3). A sugges-
tive solution is that a random mapping mechanism could be introduced to transform the phase information to a BPOF bitplane. The mapping process is protected by a cryptosystem and might prevent the proposed phase tampering attack. Another grave deficiency is the improper embedding method attributing to the easiness of extracting the hidden watermark bitplane. A conceivable countermeasure is randomly dispersing the watermark bits into multiple magnitude bitplanes, which is also controlled by an encryption mechanism. In such a manner, the proposed magnitude tampering attack might be attenuated.

Acknowledgments

The authors would like to thank Professor Jun Sang for kindly sharing source codes of the improved BPOF-based watermarking algorithm used in this paper. The authors also thank the reviewers and editor for their valuable comments. This work was supported in part by 973 Program (2011CB302204), the National Science Foundation of China for Distinguished Young Scholars (61025013), Sino-Singapore JRP (2010DFA11010), National NSF of China (61073159), and Fundamental Research Funds for the Central Universities (2009JBZ006).

References