Towards Robust Image Watermarking: Combining Content-Dependent Key, Moment Normalization, and Side-Informed Embedding

Chun-Shien Lu

Technical Report TR-IIS-03-014

Institute of Information Science, Academia Sinica

Taipei, Taiwan, ROC

E-mail: lcs@iis.sinica.edu.tw

Abstract

In digital watermarking, robustness is still a challenging problem if different sets of attacks need to be tolerated simultaneously. In this paper, we deal with this problem by using an integrated solution of side-informed embedding, moment normalization, and content-dependent watermarks. First, a new image watermarking method designed based on the concept of communications with side information is proposed. We investigate the characteristics of mean filtering in formulating new watermark embedding and extraction processes. Second, regarding resistance to geometrical attacks we do not rely on the concept of pilot signals because they are vulnerable to synchronization removal attacks. We instead use block-based watermarking and moment normalization mechanisms to recover geometrical distortions. Third, regarding resistance to the copy attack, the content-dependent watermark is employed to avoid treating an un-watermarked image as one that has been watermarked. The robustness of our approach has been verified using both the StirMark and the copy attack.

Keywords: Attack, Content-dependent watermark, Moment normalization, Robustness, Side information, Watermarking

1 Introduction

Digital watermarking [9] is now considered an efficient technology for copyright protection. This emerging area has already led to the development of numerous watermarking methods. Many requirements [4] have been recognized and evaluated in the benchmarking of watermarking systems. Among them, some parameters (e.g., fidelity, robustness) are commonly used in a variety of applications, while others (e.g., high capacity, complexity) are only employed in specific applications. Although not all requirements have to be satisfied for a specific watermarking application, robustness is still definitely important because many attacks already exist, and new attacks will appear in the future.

In general, attacks can be roughly categorized into four classes [35]: (1) removal attacks that contain non-geometrical modifications, including filtering, lossy compression, denoising, sharpening, and so on; (2) geometrical attacks that contain local/global transformations, warping, and jittering; (3) protocol attacks that are mainly composed of the copy attack and the watermark inversion attack; (4) cryptographical attacks that are related to the security of keys, such as brute fore key search and oracle attacks. Earlier robust watermarking methods could resist removal attacks [5, 19, 30]. However, past experiences indicated that resistance to removal attacks is no longer a difficulty. Consequently, geometrical attacks have recently been taken more seriously. In [31], Ruanaidh and Pun presented an RST (rotation, scaling, and translation) resilient watermarking scheme based on the Fourier-Mellin transform (FMT). The key properties they relied on are as follows: (i) the Fourier magnitude is inherently invariant to translation; and (ii) FMT provides scaling and rotation invariance along the two polar axes, respectively. The authors also noticed a weakness in that practical implementation suffers from numerical instability resulting from inverting log-polar mapping to get the watermarked image. To deal with this problem, Lin *et al.* [18] proposed an implementation inexpensive algorithm. Recently, another efficient implementation of the Fourier-Mellin transform developed by using Logarithmic Radial Harmonic Functions (LRHFs) to avoid interpolation artifacts has been proposed [8]. Despite the capability of those Fourier-Mellin transform-based watermarking methods to achieve RST invariance, their resistance to other geometrical distortions (e.g., changes of the aspect ratio and cropping) and removal attacks are limited because FMT is only RST invariant and most of the FMT information is contained in the phase instead of the magnitude part of the Fourier transformed domain.

To make a method resistant to more geometrical modifications, a certain kind of prior information (also known as the pilot signal) [24] is embedded in advance to permit the recovery of geometrical distortions at the detection stage by exploiting the known characteristics of pilot signals. In [13], Kutter was first to propose a watermarking scheme that can tolerate geometrical distortions, including RST,

change of the aspect ratio, and shearing. The key point is to embed a reference watermark, which has been arranged as a specific structural pattern in advance, for the purpose of calibration. Kutter's reference watermark is composed of nine peaks, extracted by means of an autocorrelation function and used to estimate the effects of geometrical attacks. By inverting the geometrical transformations, the hidden watermark can be recovered. The main weakness is that the other non-central eight peaks are inherently less robust to attacks. In addition, the template-based watermarking method [28], which is similar to Kutter's scheme [13], has been proposed. It inserts a template for blind registration of images that have undergone a general geometrical transformation. The template signal is designed as a specific structure with the aim to reduce the search space of template matching at the detection stage. Unfortunately, its opponent, the template removal attack [10], can successfully remove the hidden template such that geometrical transformations cannot be correctly estimated. A more powerful approach [34, 36] is to extend Kutter's scheme through block-based periodical placement of reference watermarks so that the Fourier magnitude spectrum of periodical watermarks is composed of regular peaks distributed all over the image. This particular characteristic offers the capability of recovering global/local geometrical distortions. Again, because the positioned periodical block-based pilot signals inherently reveal peaks in the transformed domain, hints are left that pirates can use to remove them. A recent paper by Manuel *et al.* [24] exhaustively analyzed pilot-based synchronization algorithms and confirmed that pilot signals are easy to be destroyed.

On the other hand, image moment normalization [1, 22] has also been proposed to recover geometrical transformations. In [2], a watermark is claimed to exist if the perturbation of moment invariants is within a small tolerance. The major disadvantages of [2] include (i) an inability to preserve fidelity, i.e., the watermarked image will create contrast variations; and (ii) an inability to tolerate any change of the aspect ratio or cropping. Recently, Bas *et al.* [3] presented a Delaunry tessellation of features points to combat some geometrical distortions. However, as noted in [14], the major drawbacks of this approach are instability and fragility of feature point extraction.

In addition to the importance of resisting geometrical attacks, the impacts of the copy attack [15] cannot be ignored because it is easy to achieve and the damage it causes is huge. This challenging problem has rarely been considered in the literature. This is because many existing robust watermarking approaches [7, 13, 34, 36] simply used denoising algorithms (e.g., Wiener filtering) to blindly extract the hidden watermark. When a denoising algorithm is used, no secret key is needed to estimate the hidden watermarks. Under these circumstances, either an owner or an un-authorized person can perform watermark estimation easily. An owner wants to extract watermarks in order to claim his

rightful ownership. On the other hand, an attacker seeks to remove/copy hidden watermarks for unauthorized use. Imagine the following scenario: a potential watermark is estimated (and then restored or decoded) from an attacked image to verify the existence of the original watermark. Meanwhile, the estimated watermark is removed to cause watermark detection to fail (the false negative problem) or added into another image to fool watermark detection (the false positive problem). We shall analyze the impacts of the above-mentioned watermark estimation-based attacks (Sec. 4) to confirm that the latter case (a copy attack, which is a kind of protocol attack) is easier to achieve than the former one (a denoising attack, which is a kind of removal attack). Having the aforementioned understanding, we should keep in mind that a robust watermarking scheme that can resist one class of attack may fail to resist another class of attack.

In view of the fact that pilot signal-based synchronization watermarking methods are vulnerable to pilot-removal attacks [24], and the copy attack [15] is easy to perform, the robustness of pilot signals and resistance to copy attacks are contradictory. However, the damage caused by the copy attack in creating protocol ambiguity is believed to be more serious than that caused by local transformations in creating the false negative problem. As a result, under the fact that the current methods are not robust to each category of attacks we would rather survive the copy attack at the expense of sacrificing resistance to local transformations. Based on the above considerations, this paper will focus on resistance to removal attacks, global geometrical transformations, and the copy attack. Our scheme is composed of three major components: (1) watermark embedding and detection; (2) image moment normalization; and (3) content-dependent watermarks. Regarding the first component, our embedding strategy is designed to be side-informed. The idea is to use a piece of prior knowledge (side information) from the cover data to facilitate the reliable extraction of watermark values blindly at the detection stage. We will give mathematical derivations and explanations on how to conceal watermarks while preserving fidelity. In addition, our embedding strategy is performed in the 8×8 DCT domain with one watermark bit concealed in each block. This is because block-based embedding is more advantageous than coefficient-based embedding in keeping information less unaffected. With this property, it is possible to tolerate geometrical distortions that remove some contents through row/column removal or general linear geometric transformation [29]. However, depending on block-based watermarking only is not sufficient to deal with other geometrical transformations, such as scaling, change of the aspect ratio, and flipping. Hence, we propose to use image moment normalization to transform an image into a canonical form before performing embedding and detection [22]. Unlike [2], our approach does not cause perceivable contrast changes during the embedding stage. The third part of our scheme

overcomes the copy attack, which has rarely been dealt with in the literature. We understand clearly that the success of the copy attack lies in the randomization of an embedded watermark and its independence from the carrier (i.e., the cover image). If the hidden watermark is closely related to the cover image, the copied watermark can carry information relevant to its carrier. Thus, our idea is to mix the hidden watermark signal with a piece of specific image-dependent information such that the practically embedded watermark is content-dependent. With this unique characteristic, even a copied watermark is forced to blend into another irrelevant carrier, so the inserted watermark will look like noise with respect to the carrier.

The remainder of this paper is organized as follows. The concept of viewing watermarking as communications with side information will be briefly introduced in Sec. 2. Based on this concept, we shall design a new block-DCT side-informed watermarking scheme in Sec. 3. In Sec. 4, the content-dependent watermark will be employed to conquer the copy attack. Finally, experimental results and concluding remarks will be, respectively, reported in Secs. 5 and 6.

2 Watermarking as Communications with Side Information

In the past few years, spread spectrum watermarking [5, 30] has been widely used for copyright protection. Its major characteristic is that it treats both the cover carrier and the watermark as noises without utilizing any possible side information or prior knowledge. This has led to the difficulty of achieving more requirements, such as transparency, robustness and blind detection, simultaneously. We have pointed out in [20] that the major drawback of previous spread spectrum-based approaches is that the prior information that can be derived from attacks [19] or detectors [21] is not properly used in the design process of an embedding strategy.

A formal formulation of viewing watermarking as communications with side information was presented in [6]. The idea behind side-informed watermarking is to utilize the prior knowledge derived from the cover carrier instead of considering it as noise. In [6], the authors proposed to maximize the probability of detection with acceptable fidelity by using a "mixing function" F, which is defined as a weighted combination between the extracted signal $\mathbf{v}$ and the watermark $\mathbf{w}$, where $\mathbf{v}$ is extracted from the cover data. The resultant mixed signal $\mathbf{s}$ is $\mathbf{s} = F(\mathbf{v}, \mathbf{w})$, which is expected to satisfy two criteria: (i) $\mathbf{s}$ should be perceptually similar to the extracted signal $\mathbf{v}$, and (ii) $\mathbf{s}$ should be highly correlated to the hidden watermark $\mathbf{w}$. Typically, the mixing function can be defined as

$$\mathbf{s} = F(\mathbf{v}, \mathbf{w}) = \mathbf{v} + \beta \cdot \mathbf{w},$$

where β is a scaling factor. In the watermark hiding process, the mixed signal $\mathbf{s}$ instead of the watermark $\mathbf{w}$ is embedded. Therefore, the signal extracted in the detection process will be a variation of $\mathbf{s}$ instead of $\mathbf{v}$. Finally, a normalized correlation is used to indicate the presence/absence of the hidden watermark $\mathbf{w}$.

3 The Proposed Image Watermarking Scheme

The proposed watermarking scheme is designed based on the concept of communications with side information [6]. In principle, our scheme can be applied to both the spatial domain and the transformed domain [22]. However, two assumptions have to be made in advance: (i) if our method is applied to the spatial domain, then we assume that the hidden watermark is actually embedded in the high-frequency components of cover data; (ii) if our method is applied to the transformed (DCT or wavelet) domain, then we assume that the cover data represents a set of selected transformed coefficients. On the other hand, the embedded watermark in this paper is a random sequence generated by a secret key or a message first converted into a binary sequence and then shuffled by means of a secret key. Either sequence is encoded using a BCH code, a kind of error correction coding (ECC), to enhance error correction. Finally, the values of the ECC encoded sequence, i.e., the watermark $\mathbf{w}$, are mapped from $\{0 \;\; 1\}$ to $\{-1 \;\; 1\}$. In what follows, we shall elaborate on our side-informed watermarking method. The maximum length of a watermark, $|\mathbf{w}|$, is equal to the number of DCT blocks in a cover image. Fig. 1 depicts the flowchart of the proposed image watermarking scheme.

3.1 Watermark Embedding

In this section, a frequency-domain watermarking method based on 8×8 block-DCT transform is proposed by casting one watermark bit into the middle frequency part of a block. Our intention is to make the proposed approach resistant to geometric distortions with content shifting. In addition, our approach does not need any help from pilot signals that are commonly used for synchronization [13, 34]. Another advantage of our approach is that if moment normalization is further used to transform an image into its canonical form before watermark embedding, then a number of geometric distortions, such as scaling or flipping, can be solved.

3.1.1 Block-wise Modulation

Let the DCT coefficients in an 8×8 block be zigzag scanned with in order from 1 to 64, as depicted in Fig. 2. We only choose some middle frequency components by skipping the DC term and the first few AC coefficients for watermarking. This is because the first few AC terms are not easy to modify sufficiently without violating fidelity. Other middle-to-high frequency coefficients are also skipped because they intrinsically tend to be eliminated by compression attacks. Therefore, this choice will play a trade-off role between transparency and robustness. Suppose the selected middle frequency components, $f_i(m_1) \sim f_i(m_2)$, in a block i (i indicates a block number) are at the locations $m_1 \sim m_2$ with $m_1 \leq m_2$ and $m_1, m_2 \in [1, 64]$. Let the mean value of the selected middle frequency coefficients of a block be $u(i)(= \frac{1}{m_2 - m_1} \sum_{k=m_1}^{k=m_2-1} f_i(k))$. These mean values will form a 1-D sequence, $\mathbf{U}$, from which our watermarking process can start.

In the embedding process, the sequence $\mathbf{U}$ is mean filtered to generate a filtered sequence $\bar{\mathbf{U}}$ with elements $\bar{u}(i)$. Let the difference between $u(i)$ and $\bar{u}(i)$ be defined as

$$d(i) = u(i) - \bar{u}(i). \tag{1}$$

These difference values, $d(i)'s$, are basically the high-frequency part of $u(i)'s$ and are treated as the extracted signal, $\mathbf{v}$, of [6] (also described in Sec. 2). At this stage, we shall develop a new methodology to embed watermarks in the transformed domain. Here, we adopt a new relationship between $u^h(i)$ and $\bar{u^h}(i)$, which are, respectively, the watermarked versions of $u(i)$ and $\bar{u}(i)$, and which can be used to accomplish watermark hiding. In the case of watermark detection for copyright protection, the embedded relationship can be used to decide on the existence of a watermark. Specifically, the above mentioned relation can be converted into a quantity, $Q(i)$, which has the same sign as its corresponding watermark value, $w(i)$. That is,

$$sgn(Q(i)) = sgn(w(i)), \tag{2}$$

where the sign function, $sgn(\cdot)$, is defined as

$$sgn(t) \quad = \quad \begin{cases} +1, & \text{if } t \geq 0, \\ -1, & \text{if } t < 0. \end{cases}$$

By means of watermark embedding, we expect to be able to impose the following relation upon watermarked images:

$$u^h(i) - \bar{u^h}(i) = sgn(w(i))|Q(i)|, \tag{3}$$

where $u^h(i)$ is either larger or smaller than $\bar{u^h}(i)$ by a magnitude $|Q(i)|$. In comparison with Eq. (1), $d(i)$ is changed to $sgn(w(i))|Q(i)|$ through embedding. In Eq. (3), $u^h(i)$ is the watermarked version of $u(i)$ obtained by means of the following modulation rule:

$$u^h(i) = u(i) + d^h(i), \tag{4}$$

with $d^h(i)$ being the modulation quantity to be determined. In this study, $d^h(i)$'s correspond to the mixed signal, $\mathbf{s}$, of [6]. Of course, $Q(i)$ has to be a constituent component of $d^h(i)$. Reorganizing Eqs. (3) and (4), $d^h(i)$ can be derived as

$$d^h(i) = sgn(w(i))|Q(i)| + \bar{u^h}(i) - u(i). \tag{5}$$

Substituting Eq. (1) into Eq. (5), the modulation quantity $d^h(i)$ can be further derived as

$$d^h(i) = sgn(w(i))|Q(i)| - d(i) + (\bar{u^h}(i) - \bar{u}(i)). \tag{6}$$

As we can see from both sides of Eq. (6), there is a recursive relation between $d^h(i)$ and $\bar{u^h}(i) - \bar{u}(i)$, which are both indexed with the superscript h to indicate hiding. More specifically, $\bar{u^h}(i)$ can be generated only after $d^h(i)$ has been determined and the embedding process has been completed (Eq. (4)), but at the same time, $d^h(i)$ has to be determined depending on $\bar{u^h}(i)$. In order to simplify the design of $d^h(i)$ in the watermark embedding process, we try to break this recursive relationship by evaluating the effect of $\bar{u^h}(i) - \bar{u}(i)$ and assuming it to be zero. Therefore, based on $\bar{u^h}(i) - \bar{u}(i) = 0$, and Eqs. (6) and (1), Eq. (4) can be re-written as

$$\begin{aligned} u^h(i) &= u(i) + d^h(i) \\ &= u(i) - d(i) + sgn(Q(i))|Q(i)| \\ &= \bar{u}(i) + sgn(Q(i))|Q(i)|. \end{aligned} \tag{7}$$

Applying mean filtering on both sides of Eq. (7), we have

$$\begin{aligned} \bar{u^h}(i) &= \bar{\bar{u}}(i) + \frac{1}{|\mathbf{n_i}|} \sum_{k \in \mathbf{n_i}} sgn(w(k))|Q(k)| \\ &= \bar{\bar{u}}(i), \end{aligned} \tag{8}$$

where the term $\frac{1}{|\mathbf{n_i}|} \sum_{k \in \mathbf{n_i}} sgn(w(k))|Q(k)|$ can be reasonably assumed to be close to zero if $Q(k)\ \forall k$ are assigned to be a constant and the distribution of $w(k)$'s is random enough, and where $\mathbf{n_i}$ denotes the neighborhood centered at i. In the following, all $Q(\cdot)$'s will be equal to the constant Q and

will be used interchangeably. Based on the assumption that $\bar{u^h}(i) - \bar{u}(i) = 0$ and the derived result $\bar{u^h}(i) = \bar{\bar{u}}(i)$, we obtain

$$\bar{u}(i) = \bar{\bar{u}}(i). \tag{9}$$

Similarly, the inverse derivation can also be established. That is, $\bar{u^h}(i) - \bar{u}(i) = 0$ if and only if $\bar{u}(i) - \bar{\bar{u}}(i) = 0$. According to the assumption that $\bar{u^h}(i) - \bar{u}(i) = 0$, $d^h(i)$ in Eq. (6) could be simplified as

$$d^h(i) = \begin{cases} +|Q(i)| - d(i), & \text{if } w(i) \geq 0, \\ -|Q(i)| - d(i), & \text{if } w(i) < 0. \end{cases} \tag{10}$$

Although there is no guarantee that Eq. (9) can always hold, our empirical results have indicated that through the use of the modulation quantity (Eq. (10)), satisfactory robustness can be achieved.

As for the transparency requirement, the modulation quantities determine the fidelity between the watermarked image and its corresponding cover image. Suppose human visual system (HVS)-based masking thresholds are used to restrict the degree of modification of DCT coefficients. Because our watermarking method is 8×8 block-based with a watermark bit cast on a finite range of middle-frequency components, we simply assume that the masking thresholds are the same for all selected coefficients within a block and are denoted as a constant JND. When JND is used to determine $d^h(i)$ (Eq. (10)), the new modulation quantity, $d^h_{HVS}(i)$, is defined as

$$d^h_{HVS}(i) = \begin{cases} MAX(d^h(i), +JND), & \text{if } d^h(i) \geq 0, \\ MAX(d^h(i), -JND), & \text{if } d^h(i) < 0, \end{cases} \tag{11}$$

where $MAX(\cdot, \cdot)$ is a maximum function. Theoretically, Eq. (10) achieves maximum robustness, while Eq. (11) plays the role of a trade-off between fidelity and robustness.

From the above description, it is clear that our idea is to increase/decrease the mean of the transformed coefficients if its corresponding watermark bit is positive/negative. This can be achieved by first eliminating the $d(i)$ term (Eq. (1)) to get $u(i) = \bar{u}(i)$. Next, a quantity $sgn(w(i))|Q(i)|$ is added to $\bar{u}(i)$ to fulfill the condition of Eq. (4). Finally, based on the assumption that Eq. (9) holds, the relation defined in Eq. (3) can be enforced. This imposed relation is the basis of the proposed method.

3.1.2 Coefficient-wise Modulation

According to Eq. (4), watermarking is actually not finished yet because only the mean value, $u(i)$, of the selected DCT coefficients has been modulated by $d^h(i)$. Every selected coefficient in a block has

to be practically modulated to accomplish embedding. Here, we propose to propagate the modulation quantity, $d^h(i)$, to all selected coefficients in a block. That is, the original coefficient, $f_i(j)$, in a block i has to be modulated as $f_i^h(j)$ for $m_1 \leq j \leq m_2$ and $m_1, m_2 \in [1, 64]$ based on the modulation rule

$$f_i^h(j) = f_i(j) + d^h(i). \tag{12}$$

By Eq. (12), we have changed from block-wise modulation to coefficient-wise modulation, and the embedding operation has now actually been accomplished. The basic idea is that, based on the property of mean filtering, it is possible to keep the modulation quantity of each selected DCT coefficient the same as that of the mean.

The principle of side-informed embedding [6] requires that a mixed signal to be highly correlated with the embedded signal. To check this condition, the correlation between our mixed signal $d^h(\cdot)$ (defined in Eq. (10)) and embedded signal $sgn(w(\cdot))Q(\cdot)$ can be calculated as

$$\begin{aligned} \sum d^h(i) \times sgn(w(i))Q(i) &= \sum (sgn(w(i))Q(i) - d(i))sgn(w(i))Q(i) \\ &= \sum Q^2 + Q \sum sgn(w(i))d(i) \\ &\approx Q^2, \end{aligned}$$

to achieve the maximum correlation because the modulation quantities, $Q(i)$'s, have been designed to be a constant, as discussed in Eq. (8), and because the term $\sum sgn(w(i))d(i)$ will naturally be a random sequence with zero mean.

3.2 Watermark Extraction

In this section, we shall analyze mean filtering facilitates watermark detection conducted without resorting to the original image and the effects of attacks on the proposed method.

As described in the previous section, each selected DCT coefficient in a block is either increased or decreased in magnitude after watermark embedding. Based on the use of mean filtering to obtain the filtered sequence $\bar{\mathbf{U}}^h$ of a watermarked image ($\bar{\mathbf{U}}^h$ is a watermarked version of $\bar{\mathbf{U}}$), we expect that the value $\bar{u}^h(i)$ of $\bar{\mathbf{U}}^h$ will consistently change when attacks are encountered. If this is true, it is sufficient to detect the hidden watermark bits through the inverse operation of the watermark modulation rule because we have increased or decreased the magnitudes of the DCT coefficients during the embedding process. One may ask: "How can the mean of those selected DCT coefficients in a block change consistently with those individual DCT coefficients even when attacks are encountered?" As we have shown in [19], the behaviors of most attacks either increase or decrease the magnitudes of

transformed coefficients. For instance, compression/sharpening tends to decrease/increase the magnitudes of transformed coefficients. These facts confirm again that mean filtering is able to move the individual DCT coefficient and the mean value of a modified (by either embedding or attacking) block toward the same polarity.

Now, we will explain how watermarks can be blindly detected when watermark modulation rules like those in Eqs. (4) and (10) are used. First, our analysis is conducted in an attack-free environment. Based on Eq. (3), $u^h(i)$ can be re-written as

$$u^h(i) = u(i) + sgn(Q(i))|Q(i)| - d(i) \quad = \quad \begin{cases} \bar{u}(i) + |Q(i)|, & \text{if } w(i) \geq 0, \\ \bar{u}(i) - |Q(i)|, & \text{if } w(i) < 0, \end{cases} \tag{13}$$

depending on the sign of its corresponding watermark value, $w(i)$. Notice that we aim to have $sgn(Q(i)) = sgn(w(i))$, as expressed in Eq. (2). Once all the $u^h(i)$ values have been yielded for all blocks, they are mean filtered to obtain $\bar{u^h}(i)$. However, based on the assumption that $\bar{u^h}(i) - \bar{u}(i) = 0$, made in the previous section, the detected watermark bit can be blindly determined according to Eqs. (3) and (13) based on an attack-free environment:

$$w^e(i) = sgn(u^h(i) - \bar{u^h}(i)) = sgn(u^h(i) - \bar{u}(i)) = sgn(Q(i)) \quad = \quad \begin{cases} +1, & \text{if } Q(i) \geq 0, \\ -1, & \text{if } Q(i) < 0. \end{cases} \tag{14}$$

The above detection result is exactly consistent with what we have designed in Eq. (3). Generally speaking, the empirical results and theoretic results are very close to each other even under the assumption that $\bar{u^h}(i) - \bar{u}(i) = 0$.

Now, we will discuss the problem of watermark detection in an attack-prone environment. Suppose the effects caused by attacks include channel fading, **c**, and channel noise, **a**. Under these circumstances, the watermarked DCT coefficient in Eq. (12) becomes

$$f_i^a(j) = c_i(j) f_i^h(j) + a_i(j). \tag{15}$$

From Eq. (15), its mean filtered version can be written as

$$u^a(i) = C(i) u^h(i) + A(i), \tag{16}$$

where $C(i)$ and $A(i)$ are, respectively, the mean of $c_i(j)$'s and $a_i(j)$'s in a block i. After applying mean filtering to $u^a(i)$'s, we obtain

$$\bar{u^a}(i) = \bar{C}(i) \bar{u^h}(i) + \bar{A}(i). \tag{17}$$

Therefore, the detected watermark bit is blindly determined (as previously indicated in Eq. (14)) by subtracting Eq. (17) from Eq. (16) and by taking the sign as

$$\begin{aligned} w^e(i) &= sgn(u^a(i) - \bar{u^a}(i)) \\ &= sgn((C(i)u^h(i) - \bar{C}(i)\bar{u^h}(i)) + (A(i) - \bar{A}(i))) \\ &= sgn(\alpha Q(i) + (A(i) - \bar{A}(i))), \end{aligned} \tag{18}$$

where $C(i)$ and $\bar{C}(i)$ are assumed to be equal for simplicity; i.e., the fading effects act equally on the whole image. In Eq. (18), the first term in the sgn function is the same as Eq. (14), whereas the second term represents the noise term. As we can see from Eq. (18), if α is small enough (i.e., the attack is strong) such that $|A(i) - \bar{A}(i)| > |\alpha Q(i)|$ and $sgn(\alpha Q(i)) \neq sgn(A(i) - \bar{A}(i))$, then the effect of watermarking will disappear. This means that the embedded watermark has been destroyed or removed. Hence, this explains why the modulation quantity and the channel parameters both play important but contradictory roles with respect to the requirement of robustness. In either of the above two attacking environments, the normalized correlation, ρ, between $\mathbf{w}$ and $\mathbf{w^e}$, defined as

$$\rho = \mathbf{w} \cdot \mathbf{w^e} = \frac{1}{|\mathbf{w}|} \sum_{k=1}^{|\mathbf{w}|} sgn(w(k)) sgn(w^e(k)), \tag{19}$$

is used to indicate the presence/absence of a watermark, where sgn is the sign function defined in Eq. (3).

As mentioned previously, the fading effect is, at present, ignored in the proposed watermarking approach. Usually, significant fading leads to severe quality degradation of the cover data, which can no longer be authentic as discussed in [17, 23]. In particular, an image cannot be considered perceptually acceptable if the applied compression ratio is large. In addition, an image, even one considered to not be authentic, can still be used by an unauthorized person. The above discussion reflects the major difference between copyright protection and content authentication. However, due to the fact that a piece of stego data will be useless if its quality is severely degraded by an attack, attacks should only be imposed within the constraint of acceptable degradation (i.e., the fading effect cannot be large). Therefore, our scheme can maintain robustness to some extent as long as the introduced fading effect is not unlimited.

3.3 Resistance to Geometric Distortions Using Block-based Embedding and Moment Normalization

In this section, we will describe the capability of our scheme in resisting geometric distortions without resorting to pilot signals or reference watermarks [13, 34] based on two aspects, i.e., block-based embedding and moment normalization. Recall that the proposed scheme is performed in the block-DCT domain. The property of block transformation can be relied to tolerate pixel/line deleting or shifting such as row/column removing, general linear geometric transformation, and change of the aspect ratio. This is mainly due to the fact that only one bit is inserted into a block, and that block-wise embedding is more stable than pixel-wise embedding (i.e., the block's characteristic is more robust than the pixel's characteristic). However, depending only on block DCT is not sufficient because the size of an image may be changed significantly by scaling. In order to compensate for the effect of global transformations, moment normalization [1, 22] has been employed to transform an image into a canonical form before embedding and detection are performed. In other words, watermarks are embedded/detected into/from the canonical form of a cover image. More detailed operations of image normalization can be found in [1, 27, 32]. By taking into account the features of block-based watermarking and moment normalization, our method is able to resist numerous geometric distortions, including scaling, change of the aspect ratio, line removing, general linear geometric transformation, flipping, and rotation without cropping [29]. Compared with moment invariants-based watermarking [2], our scheme can resist more geometrical distortions. This is mainly due to the difference in the developed embedding strategies. However, we should note that the proposed scheme still does not take local geometrical distortions into consideration. As stated in Sec. 1, our purpose is to overcome the copy attack (described in Sec. 4) with higher priority. To recover local transformations, the diversity property [12, 34] is an alternative that should be further incorporated.

4 Resistance to Copy Attack Using Content-Dependent Watermark

While numerous methods have been claimed to be robust, they are actually only robust from one perspective. From another perspective, they are extremely fragile. One vital challenge to robustness is the copy attack [15], which has been developed to create the false positive problem; i.e., one can successfully detect a watermark from an unwatermarked image. The copy attack is operated as follows: (i) a watermark is first predicted from a stego image; (ii) the predicted watermark is added into a target image to create a counterfeit watermarked image; and (iii) from the counterfeit image, a watermark

can be detected that wrongly claims rightful ownership. In the light of this simple but strong attack, we have an open question: "Does *successful* prediction of a watermark also imply that removal and copy of a watermark can be done successfully?" [21]. In this paper, we first conduct analyses to show the significance of the copy attack and then present a solution to deal with this problem.

4.1 Watermark-Estimation Attacks

In this section, two kinds of watermark-estimation attacks, i.e., the denoising attack and the copy attack, will be discussed. From our analyses, the ease of performing a copy attack will be shown. Without loss of generality, the decision on a watermark's existence will be based on correlation, as defined in Eq. (19). Let X, X^w, Z, and Z^w be denoted as the original image, watermarked image, faked original image, and faked watermarked image, respectively. Among them, X^w is generated from X through an embedding process, and Z^w is generated from the combination of Z and a watermark estimated based on X^w. Here, watermark estimation is assumed to be conducted by means of a denoising process [15, 24, 33].

Let $\mathbf{w}$ be a watermark to be hidden in X, and let $\mathbf{w^x}$ be an estimated watermark obtained by denoising X^w. For the purpose of watermark removal [33], $\mathbf{w^x}$ will be subtracted from X^w (by omitting some scaling factors for simplicity) to form an attacked image D, i.e.,

$$D = X^w - \mathbf{w^x}.$$

In the watermark detection process, a watermark, $\mathbf{w^z}$, is extracted from D and correlated with $\mathbf{w}$. A watermark does not exist if ρ is smaller than a threshold T ($0 \leq T \leq 1$). Under this circumstance, if denoising-based watermark removal is expected to succeed, then $\rho < T$ must hold. This result indicates that the ratios of the correctly (C_w) and wrongly (NC_w) decoded watermark bits should, respectively, satisfy

$$C_w \leq \frac{1+T}{2}$$

and

$$NC_w \geq \frac{1-T}{2}, \tag{20}$$

where $C_w + NC_w = 1$ and NC_w corresponds to the bit error rate (BER). Based on false analyses [25], if we would like to have a false positive probability of less than 10^{-6}, then the threshold T should be set to be 0.15. As a consequence, the above analyses show that an efficient watermark removal attack should be able to destroy or estimate *most* watermark bits since T is usually small. In fact, the actual number has been specified in Eq. (20).

As for the copy attack [15], the estimated watermark $\mathbf{w^x}$ is copied and added to the target image Z (again, omitting some scaling factors for simplicity) to form a counterfeit image Z^w, i.e.,

$$Z^w = Z + \mathbf{w^x}. \tag{21}$$

In the watermark detection process, a watermark, $\mathbf{w^z}$, is extracted from Z^w and correlated with $\mathbf{w}$. A watermark exists if ρ is larger than or equal to a pre-determined threshold T. Under this circumstance, if a copy attack is expected to succeed, then $\rho \geq T$ must hold. This implies that the ratio of the correctly decoded watermark bits only needs to be at least increased from $\frac{1}{2}$ (due to the randomness of an arbitrary image, Z) to $\frac{1+T}{2}$. Actually, the amount of increase, ξ^{copy}, only needs to satisfy

$$\xi^{copy} \geq \frac{1+T}{2} - \frac{1}{2} = \frac{T}{2}. \tag{22}$$

Comparing Eqs. (20) and (22), we find that a copy attack is easier to perform successfully than a denoising attack because $\frac{1-T}{2} >> \frac{T}{2}$ holds since T is usually a smaller number [25].

4.2 Content-Dependent Watermark

From the above analyses, we have found that the success of the copy attack is essentially due to the independence of the watermark from the carrier. To deal with this problem, the watermark should carry information relevant to the cover image. Meanwhile, the content-dependent information must also be robust to digital processing in order not to affect the detection results. In this paper, since one watermark bit is inserted into an 8×8 block, we need to find a piece of representative but robust information for each block. Next, the content-dependent information must be incorporated into the watermark signal to generate a content-dependent watermark. Here, a representative feature in a DCT block is defined by the relationship between two AC coefficients:

$$CDK(i) = \begin{cases} +1, & \text{if } |f_i(p_1)| - |f_i(p_2)| \geq 0, \\ -1, & \text{otherwise,} \end{cases} \tag{23}$$

where $CDK(i)$ is a value in the content-dependent feature sequence (also called the content-dependent key) **CDK**, and $f_i(p_1)$ and $f_i(p_2)$ are two AC coefficients at positions p_1 and p_2 in block i. The DC coefficient will not be selected because it is positive and, thus, not random. In addition, the two selected AC coefficients should be at lower frequencies because high-frequency coefficients are vulnerable to attacks. In this paper, $p_1 = 2$ and $p_2 = 3$ are used based on the 64 available frequency subbands (notice that the lowest frequency is at the 1st subband), as shown in Fig. 2.

When the **CDK** is extracted, it is mixed with the watermark $\mathbf{w}$ to generate the content-dependent watermark (**CDW**) as

$$CDW(i) = w(i) \times CDK(i), \quad \forall i. \tag{24}$$

The sequence **CDW** is what we will embed into a cover image.

Next, we proceed to show why the presented content-dependent watermark can prevent the copy attack. Let $\mathbf{CDK_X}$ and $\mathbf{CDK_Z}$ denote the content-dependent keys generated from two different images, X and Z, respectively. In addition, let $\mathbf{CDW_X}$ denote the content-dependent watermark to be hidden into the cover image X. As described previously, let the denoising-based watermark estimated from X^w be $\mathbf{w^x}$, which will contain partial information from $\mathbf{CDW_X}$. By directing the copy attack at the target image Z, we can get the counterfeit watermarked image Z^w as defined in Eq. (21). Later, in the detection process, the estimated content-dependent watermark bit, $w^z(i)$, at block i from Z^w will be

$$w^z(i) = (\eta(i) \times CDW_X(i) + N(i)), \tag{25}$$

where the variable factor $\eta(\cdot)$ indicates the estimated watermark energy in each block, and $N(\cdot)$ indicates the noise value of the noise sequence $\mathbf{N}$ (which is irrelevant to watermarks) generated by denoising Z^w. Under the assumption that denoising is an efficient way to estimate watermarks [13, 34, 36], $\eta(\cdot) > |N(\cdot)|$ can undoubtedly hold. Given Eqs. (24) and (25), the correlation (defined in Eq. (19)) between $\mathbf{CDW_Z}$ and $\mathbf{w^z}$ is easily derived as

$$\begin{aligned}
\rho &= \frac{1}{|\mathbf{w}|}\sum_{i=1}^{|\mathbf{w}|} sgn(CDW_Z(i))sgn(w^z(i)) \\
&= \frac{1}{|\mathbf{w}|}\sum_{i=1}^{|\mathbf{w}|} sgn(w(i))sgn(CDK_Z(i))sgn(N(i)+\eta(i)CDK_X(i)w(i)) \\
&\approx \frac{1}{|\mathbf{w}|}\sum_{i=1}^{|\mathbf{w}|} sgn^2(w(i))sgn(CDK_X(i))sgn(CDK_Z(i)),
\end{aligned} \tag{26}$$

where the term $\eta(i)CDK_X(i)w(i)$ in the 2nd row can be assumed to dominate $N(i)$ based on the fact that denoising is efficient for estimating watermarks. When the copy attack is performed, it is not hard to check that the correlation (Eq. (26)) derived from a counterfeit image will be approximately zero since any two random content-dependent keys, $\mathbf{CDK_X}$ and $\mathbf{CDK_Z}$, are nearly orthogonal. On the other hand, if watermark detection is conducted on a watermarked image, then the obtained ρ $(=\frac{1}{|\mathbf{w}|}\sum_{i=1}^{|\mathbf{w}|} sgn^2(w(i)))$ can be perfectly high due to $\mathbf{CDK_X} = \mathbf{CDK_Z}$.

4.3 Remarks on Content-Dependent Watermarks

In fact, Eq. (23) defines the relationship between a pair of transformed coefficients. This feature is similar to either the inter-block relationship of DCT coefficients [17] or the parent-child relationship of wavelet coefficients [23]. Both have been successfully applied in image authentication [17, 23] and can be extended to fingerprinting [26]. In content authentication, both robustness and fragility are required to tolerate incidental manipulations and reflect malicious modifications. However, considering the problem of distinguishing between two images only robustness is needed. As a result, the pair of coefficients must be carefully selected to achieve robustness. To this end, low-frequency coefficients are used, as suggested in the previous paragraph. In addition, we have to emphasize that the robustness with respect to the relationship between a pair of coefficients is not totally due to the low frequency characteristic. On the contrary, the relationship defined in Eq. (23) plays a key role in resistance against digital processing. Since this topic is beyond the scope of this paper, readers can refer to relevant researches [17, 23, 26] for more details.

5 Experimental Results

We conducted a series of experiments to evaluate the performance of the proposed robust watermarking scheme. Each hidden watermark herein was dependent upon the cover image. In this study, the middle-frequency DCT coefficients within a block selected for embedding were located in the $4th \sim 10th$ subbands, as shown in Fig. 2. The BCH(63,7) code was used for error correction coding of the embedded content-dependent watermark. The support of 1-D mean filtering used in the watermark embedding and detection processes is 15. Our experiments were divided into three parts: in the first part of experiments we tried to demonstrate the relationship between the acceptable fidelity and the accuracy of blind detection for watermarked but not attacked images; the second part of experiments were conducted to test robustness against the StirMark attack [29]; and the third part of experiments were conducted to test robustness against the copy attack [15], which is a kind of protocol attack.

In our experiments, ten varieties of color cover images, as shown in gray-scale in Fig. 3, were used for performance evaluation. Among them, the first six images contained large homogeneous areas, while the remaining four images contained mostly texture/edge areas. Our both watermark embedding and detection operations were conducted in the Y channel of the $YCbCr$ color space.

5.1 Experiments–Part I: Fidelity vs. Blind Detection

For the cover images depicted in Fig. 3, we first examined the fidelity and accuracy of blind detection for watermarked (but not attacked) images. Generally speaking, the modulation quantity Q can be set arbitrarily large but it must be bounded by a masking threshold JND as Eq. (11) shows. In our scheme, the watermark quantity, Q, is fixed to facilitate analyses of embedding and detection, as derived in Eq. (8). Results for PSNR and the correlation ρ for the watermarked versions of Fig. 3 are given in Table 1. We can see that some detected correlation values were not perfectly lossless. The reason is mainly due to the constraint of embedding (Eq. (11)) that more or less affects the watermark quantities that are expected to be added. Fortunately, the resultant bit error rate did not have an apparent impact on robustness. For illustration of perceptual comparisons, one smoothing image (Splash) and one complex image (Tank) were selected for fidelity checking. These two cover images and their corresponding stego versions are shown in Fig. 4. No differences were subjectively perceived.

5.2 Experiments–Part II: Robustness vs. Removal and Geometrical Attacks

In the second part of experiments, the StirMark benchmark [29] was used to generate different kinds of attacks, excluding those with *large* cropping effects. The tested attacks included signal enhancement (sharpening, median filtering, Gaussian filtering, and FMLR filtering), JPEG compression, and geometrical distortions (change of the aspect ratio, flipping, general linear geometric transformation, row/column removing, scaling, rotation using small degrees, slight cropping, and random bending). For more details about the parameters of StirMark attacks, please refers to [29].

The detection result (normalized correlation, ρ) was compared with a threshold T to determine the existence of a watermark. If we wish to have a false positive probability as low as 10^{-6}, T should be set to be 0.15 [25]. With this setting, if ρ is larger than T, then a detection is regarded to be successful. In this study, a score is measured for each attack to indicate the degree of robustness. It is defined to be $\frac{number\ of\ successful\ detections}{number\ of\ different\ sets\ of\ parameters\ applied\ to\ an\ attack}$. As a result, a larger score implies higher robustness against a specific attack. Table 2 (except for the last row) depicts the robustness verification results under StirMark. It can be observed that our scheme tolerated flipping, change of the aspect ratio, and line removing very well. As for scaling, our scheme fail to resist some transformed images scaled down to quarter sizes since most of the information was lost. It is also interesting to note that even though no pilot signals (reference watermarks) were used to recover synchronization, our scheme offers resilience to rotation performed with a small degree. We conjecture that resistance to rotation with

such a small degree may be due to the use of block-based embedding to account for slight changes. Resistance to general linear geometric transformation achieved modest robustness, while resistance to cropping, shearing, and random bending was unacceptable because these are local transformations, but moment normalization can only deal with global transformations. Besides, most removal attacks could be resisted quite well with a few exceptions, including JPEG compression with small quality factors (e.g., 10%) and median filtering with a large window. The images that revealed the above exceptions mainly had large homogeneous areas (Figs. 3(a)$\sim$(e)). The reasons are twofold: (i) homogeneous regions can only be embedded with less watermark energy; and (ii) removing watermarks by means of denoising is quite easy in homogeneous regions.

In addition to evaluation of robustness on single attacks (as stated in the above), the performance of our scheme verified on combined attacks is shown in Table 3. Since our scheme has not really compensated for the effects of local transformations, they will be ignored in this test. The used combined attacks are generated by integrating JPEG2000 compression [11] with 1 bit/pixel (bpp) and the SitrMark 3.1. By comparing Table 2 and Table 3, we can observe that the capabilities of robustness are mainly reduced for the combined attacks, i.e., JPEG2000 with 1 bpp+JPEG with lower quality factors, on smoothing images. Moreover, resistance to attacks for the image shown in Fig. 3(e), which looks relatively smoothing in both foreground and background, has been reduced remarkably. As for the other combined attacks (e.g., JPEG2000 + global transformations), our scheme still works stably, i.e., recovery of global geometrical transformations was not affected by additionally applied JPEG2000 compression with 1 bpp.

Finally, we also conducted an experiment to demonstrate the immunity of our scheme to interpolation and quantization errors caused by rotations. Each stego image was rotated with small rotation angles $\pm\theta_1$ and large rotation angles θ_2, respectively, where $\theta_1 = 0.25 + 0.25q_1$ $(0 \leq q_1 \leq 3)$ and $\theta_1 = 5 + 5q_2$ $(0 \leq q_2 \leq 70)$, to generate 40 rotated images. Among them, small rotation angles have been particularly concerned in [24] because they tended to result in imperfect recovery of synchronization such that the detection performance was degraded. Here, a total of 400 rotated images were generated from the watermarked versions of the ten cover images (Fig. 3). Our watermark detection results showed that almost all the correlation values generated from all the rotated images were larger than 0.4 and were significantly larger than $T(= 0.15)$, which implies that our scheme is, indeed, immune to recovery errors.

5.3 Experiments–Part III: Robustness vs. Copy Attack

The final part of experiments was conducted to verify the capability of our scheme in resisting the copy attack. The Lee's Wiener filter [16] was used to performed denoising-based watermark estimation. We shall take the watermarked version of the Lenna image as an example for the purpose of illustration. In Fig. 5, subfigures (a) and (b) show the watermarks estimated from the image shown in Fig. 4(b) with two different sets of denoising parameters, respectively. Denoising parameters are closely related to the strength of a denoised noise, which corresponds to the energy of an estimated watermark. We intended to use different watermark energies to observe their impacts on the resultant copy attacks. The cover images, shown in Fig. 3, were used as target images for producing counterfeit images. Images with homogeneous areas can easily reveal copied watermarks remarkably, thus increasing the challenge of preventing the copy attack. Examples of counterfeit images that contain Lenna watermarks are shown in Figs. 5(c)$\sim$(f). In these subfigures, the estimated watermarks, Figs. 5(a) and (b), have been, respectively, added into (c), (e), and (d), (f). As expected, Figs. 5(d) and (f) contains stronger copied signals than Figs. 5(c) and (e), respectively. Furthermore, we also find that the copied watermarks clearly appeared in the fake images. Even though they have lost commercial value, we allow these apparent artifacts for the purpose of strict robustness test. After watermark detection was performed using our scheme, the correlation values fell into the range $[-0.132 \;\; 0.116]$ and were lower than the detection threshold T, which implied that the copy attack failed to create protocol ambiguity (the results are given in the last row of Table 2). Among the obtained results, the maximum correlation was found for the image shown in Fig. 5(d). We inferred that the main reason was that the copied watermark was strong enough to readily manifest its fake effect on such a homogeneous image. In short, we can conclude that the experimental results consist with the analyzed result (Sec. 4) quite well.

6 Concluding Remarks

Finding a robust watermarking scheme was the main focus of this study. Therefore, removal attacks, geometrical attacks, and protocol attacks were investigated. In this paper, we have proposed a new block DCT-based watermarking method based on the concept of communications with side information. Our focus has been to mathematically formulate the problem of watermark embedding and detection by investigating the characteristics of side information extracted from cover data by means of mean filtering. This embedding strategy has achieved partial resistance to geometrical transfor-

mations. By further introducing image moment normalization, more geometrical distortions could be resisted. Although our approach, similar to existing methods, cannot resist all geometrical transformations, one novel contribution of the proposed scheme is its capability to cope with the copy attack by introducing the concept of content-dependent watermark. Overall, the robustness of our method has been verified by applying a number of attacks, including StirMark 3.1 and the copy attack, to several varieties of images.

We must note that the major limitation of this approach is its fragility in resisting those attacks involving cropping, which have been popularly dealt with using pilot-based synchronization algorithms. However, as has been discussed in [24], pilot signals have their own fatal weakness in that they are easily destroyed. Consequently, enhancing the robustness of pilot signals is a worthwhile direction for further research.

Acknowledgment: This paper was supported in part by the National Science Council under NSC grant 89-2750-P-001-022-5. The author thanks Prof. Yu for providing the program for image moment normalization.

References

[1] M. Alghoniemy and A. H. Tewfik, "Geometric Distortion Correction Through Image Normalization," *Proc. Int. Conf. on Multimedia and Expo*, 2000.

[2] M. Alghoniemy and A. H. Tewfik, "Image Watermarking by Moment Invariants," *Proc. IEEE Int. Conf. Image Processing*, Vancouver, Canada, Vol. II, pp. 73-76, 2000.

[3] P. Bas, J. M. Chassery, and B. Macq, "Geometrically Invariant Watermarking Using Feature Points," *IEEE Trans. on Image Processing*, Vol. 11, No. 9, pp. 1014 1028, 2002.

[4] "Benchmark Metrics and Parameters," Certimark project no. IST-1999-10987.

[5] I. J. Cox, J. Kilian, F. T. Leighton, and T. Shamoon, "Secure Spread Spectrum Watermarking for Multimedia," *IEEE Trans. on Image Processing*, Vol. 6, pp. 1673-1687, 1997.

[6] I. J. Cox, M. L. Miller, and A. McKellips, "Watermarking as Communications with Side Information," *Proc. of the IEEE*, Vol. 87, No. 7, pp. 1127-1141, 1999.

[7] G. Depovere, T. Kalker, and J. P. Linnartz, "Improved Watermark Detection Using Filtering before Correlation," *Proc. IEEE Int. Conf. On Image Processing*, Vol. I, pp. 430-434, 1998.

[8] P. A. Fletcher and K. G. Larkin, "Direct embedding and Detection of RST Invariant Watermarks," *Int. Workshop on Information Hiding*, LNCS 2578, pp. 129 144, 2002.

[9] F. Hartung and M. Kutter, "Multimedia Watermarking Techniques," *Proceedings of the IEEE*, Vol. 87, pp. 1079-1107, 1999.

[10] A. Herrigel, S. Voloshynovskiy, and Y. Rytsar, "The Watermark Template Attack," *Proc. SPIE: Security and Watermarking of Multimedia Contents II*, Vol. 4314, USA, 2001.

[11] "http://jj2000.epfl.ch/", *An Implementation of the JPEG2000 Standard in Java*.

[12] D. Kundur and D. Hatzinakos, "Diversity and Attack Characterization for Improved Robust Watermarking," *IEEE Trans. on Signal Processing*, Vol. 29, No. 10, pp. 2383-2396, 2001.

[13] M. Kutter, "Watermarking Resistant to Translation, Rotation, and Scaling," *Proc. SPIE Int. Symp. on Voice, Video, and Data Communication*, 1998.

[14] M. Kutter, S. K. Bhattacharjee, and T. Ebrahimi, "Toward Second Generation Watermarking Schemes," *Proc. IEEE Int. Conf. on Image Processing*, Vol. I, pp. 320 323, 1999.

[15] M. Kutter, S. Voloshynovskiy, and A. Herrigel, "The Watermark Copy Attack," *Proc. SPIE: Security and Watermarking of Multimedia Contents II*, Vol. 3971, San Jose, CA, USA, 2000.

[16] J. S. Lee, "Digital Image Enhancement and Noise Filtering by Use of Local Statistics", *IEEE Trans. on Pattern Anal. and Machine Intell.*, Vol. 2, No. 2, pp. 165-168, 1980.

[17] 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 Technology*, Vol. 11, No. 2, pp. 153-168, 2001.

[18] C.-Y. Lin, M. Wu, Y. M. Lui, J. A. Bloom, M. L. Miller, and I. J. Cox, "Rotation, Scale, and Translation Resilient Public Watermarking for Images," *IEEE Trans. on Image Processing*, Vol. 10, No. 5, pp. 767-782, 2001.

[19] C. S. Lu, S. K. Huang, C. J. Sze, and H. Y. Mark Liao, "Cocktail Watermarking for Digital Image Protection," *IEEE Trans. on Multimedia*, Vol. 2, No. 4, pp. 209-224, 2000.

[20] C. S. Lu and H. Y. Mark Liao, "Digital Watermarking: A Communications with Side Information Perspective," *Proc. IEEE Pacific-Rim Conf. on Multimedia*, Beijing, China, LNCS 2195, pp. 927-932, 2001.

[21] C. S. Lu, H. Y. Mark Liao, and M. Kutter, " Denoising and Copy Attacks Resilient Watermarking by Exploiting Knowledge at Detector," *IEEE Trans. on Image Processing*, Vol. 11, No. 3, pp. 280-292, 2002.

[22] C. S. Lu, "Block DCT-based Robust Watermarking Using Side Information Extracted by Mean Filtering," *Proc. 16th IAPR Int. Conf. on Pattern Recognition*, Quebec, Canada, Vol. II, 2002.

[23] C. S. Lu and H. Y. Mark Liao, "Structural Digital Signature for Image Authentication: An Incidental Distortion Resistant Scheme", to appear in *IEEE Trans. on Multimedia*, Vol. 5, No. 2, 2003.

[24] A. R. Manuel and P. G. Fernando, "Analysis of Pilot-based Synchronization Algorithms for Watermarking of Still Images," *Signal Processing: Image Communication*, Vol. 17, pp. 611-633, 2002.

[25] M. L. Miller and J. A. Bloom, "Computing the Probability of False Watermark Detection," *3rd Int. Information Hiding Workshop*, LNCS 1768, pp. 146-158, Dresden, Germany, 1999.

[26] IEEE Int. Workshop on Multimedia Signal Processing, special session on Media Recognition, 2002.

[27] S. C. Pei and C. N. Lin, "Image Normalization for Pattern Recognition," *Image and Vision Computing*, Vol. 13, No. 10, pp. 711-723, 1995.

[28] S. Pereira and T. Pun, "Robust Template Matching for Affine Resistant Image Watermarks", *IEEE Trans. on Image Processing*, Vol. 9, pp. 1123-1129, 2000.

[29] F. Petitcolas, R. J. Anderson, and M. G. Kuhn, "Attacks on Copyright Marking Systems," *Second Workshop on Information Hiding*, USA, pp. 218-238, 1998.

[30] C. I. Podilchuk and W. Zeng, "Image-Adaptive Watermarking Using Visual Models," *IEEE Journal on Selected Areas in Communications*, Vol. 16, pp. 525-539, 1998.

[31] J. J. K. O'Ruanaidh and T. Pun, "Rotation, Scale, and Translation Invariant Spread Spectrum Digital Image Watermarking," *Signal Processing*, Vol. 66, No. 3, pp. 303 317, 1998.

[32] D. Shen and H. Horace, "Generalized Affine Invariant Image Normalization," *IEEE Trans. Pattern Anal. and Machine Intelligence*, Vol. 19, No. 5, pp. 431-440, 1997.

[33] S. Voloshynovskiy, S. Pereira, A. Herrigel, N. Baumgartner, and T. Pun, "Generalized Watermarking Attack Based on Watermark Estimation and Perceptual Remodulation," *Proc. SPIE: Security and Watermarking of Multimedia Contents II*, Vol. 3971, USA, 2000.

[34] S. Voloshynovskiy, F. Deguillaume, S. Pereira, and T. Pun, "Optimal Adaptive Diversity Watermarking with Channel State Estimation," *Proc. SPIE: Security and Watermarking of Multimedia Contents III*, Vol. 4314, USA, 2001.

[35] S. Voloshynovskiy, S. Pereira, V. Iquise, and T. Pun, "Attack Modelling: Towards a Second Generation Watermarking Benchmark," *Signal Processing*, Vol. 81, pp. 1177-1214, 2001.

[36] S. Voloshynovskiy, F. Deguillaume, and T. Pun, "Multibit Digital Watermarking Robust against Local Nonlinear Geometrical Distortions," *Proc. IEEE Int. Conf. on Image Processing*, pp. 999 1002, 2001.

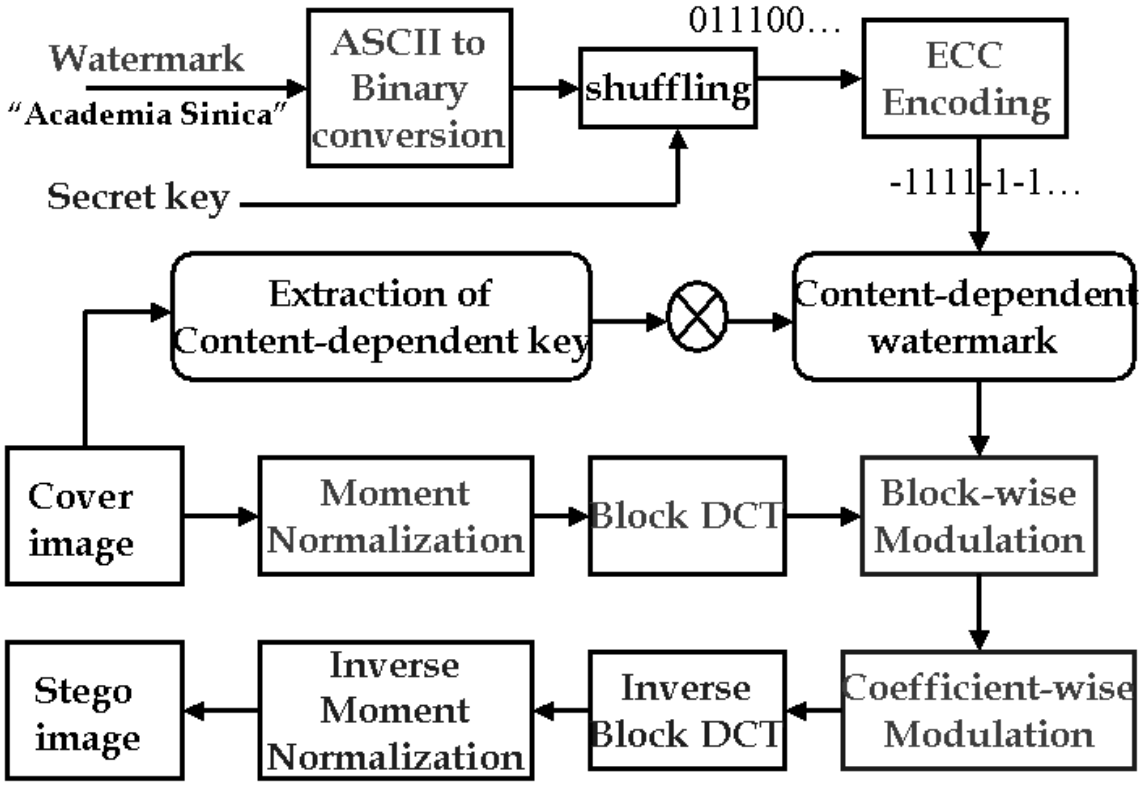


Figure 1: Diagram of the proposed block-DCT side-informed image embedding scheme. Our blind detection process is basically an inverse embedding operation.

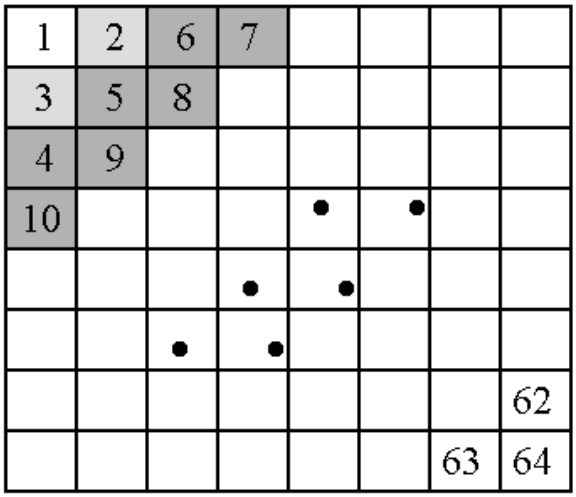
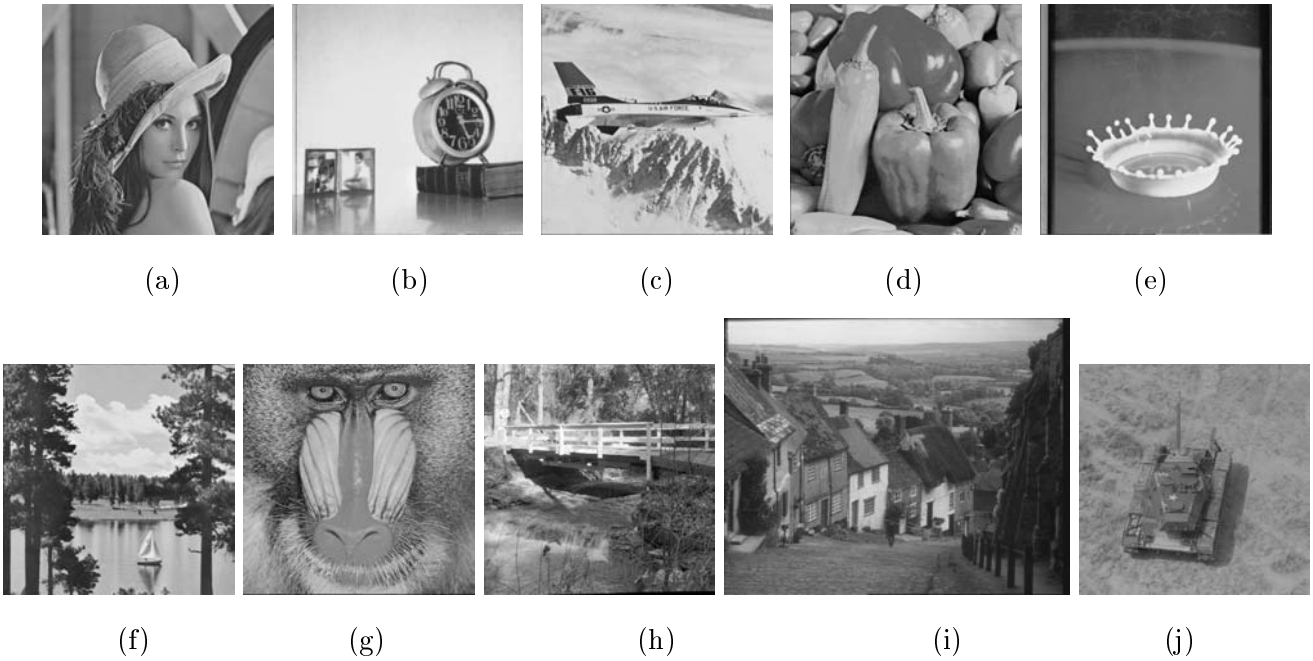


Figure 2: 8×8 DCT block: grids $4 \sim 10$ indicate the middle-frequency subbands selected for embedding; grids 2 and 3 are used to generate the content-dependent key (**CDK**).



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

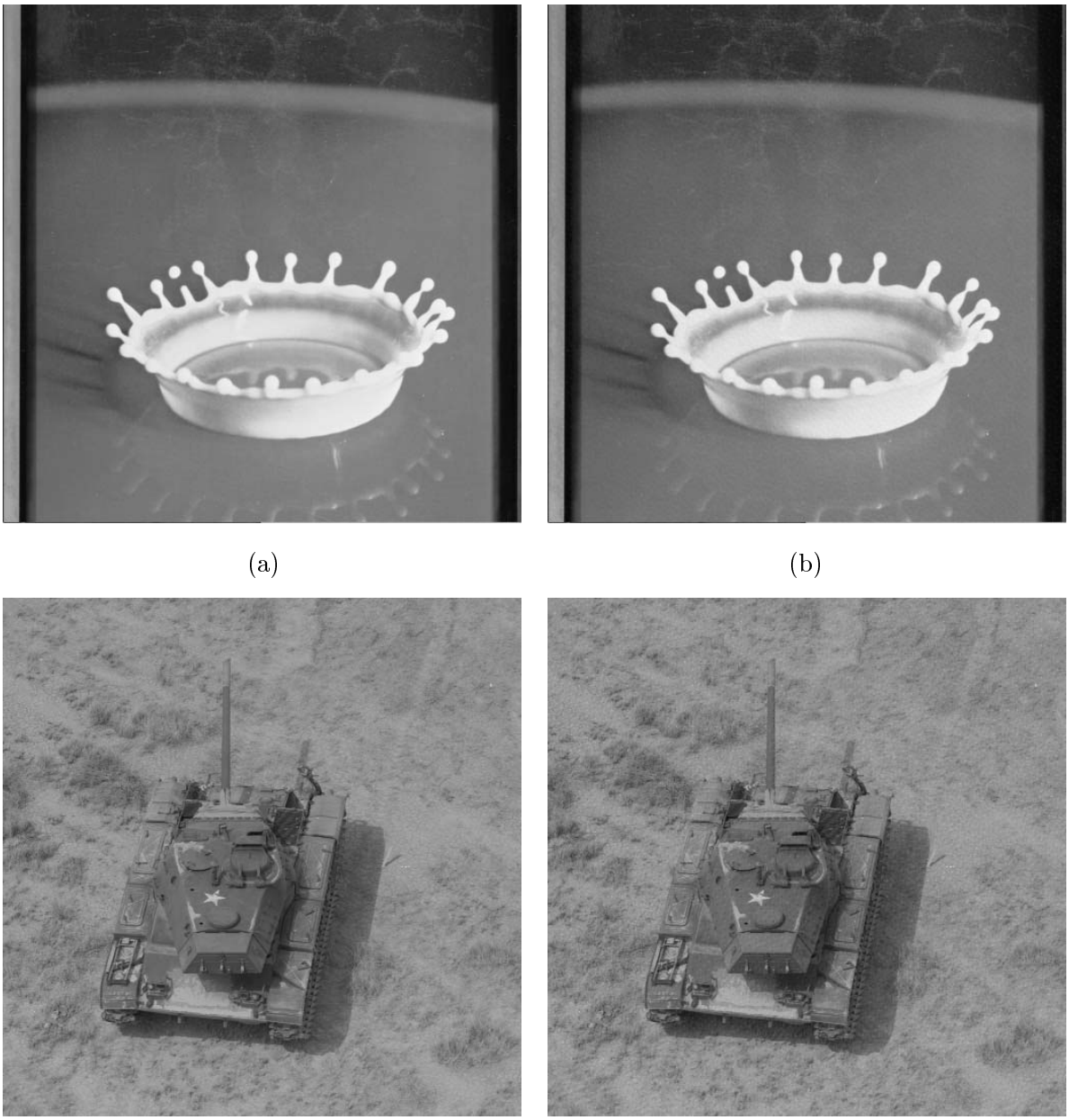
(f) (g) (h) (i) (j)

Figure 3: Tested cover images: (a)∼(f) are considered to be smoother than (g)∼(j).

Table 1: **Transparency and blind detection results of stego images w.r.t. the cover images shown in Fig. 3**

.

images	3(a)	3(b)	3(c)	3(d)	3(e)	3(f)	3(g)	3(h)	3(i)	3(j)
PSNR	43.94	43.90	43.99	40.47	43.86	44.03	36.86	36.89	38.67	38.65
ρ	0.93	1.00	1.00	1.00	1.00	0.77	0.90	0.89	0.77	1.00



(a) (b)

(c) (d)

Figure 4: Perceptual comparison of unwatermarked and watermarked images: (a) and (c) are cover images; (b) and (d) are stego images.

Table 2: **Verification of robustness w.r.t. removal, geometrical, and copy attacks. Among the attacks, the removal and geometrical attacks were generated by means of StirMark 3.1 [29]. Each attack's name is followed by a digit, which indicates the number of times that the attack was performed with different parameters.**

Attacks	3(a)	3(b)	3(c)	3(d)	3(e)	3(f)	3(g)	3(h)	3(i)	3(j)
MF (3)	1	1	1	0.67	0.67	0.67	1	1	1	1
GF (1)	1	1	1	1	1	1	1	1	1	1
FMLR (1)	1	1	0	1	0	1	1	1	1	1
Sharpening (1)	1	1	1	1	1	1	1	1	1	1
Color reduction (1)	1	1	1	1	1	1	1	1	1	1
JPEG (12)	0.83	0.5	0.83	1	0.5	0.92	1	1	0.92	0.83
Flipping (1)	1	1	1	1	1	1	1	1	1	1
Scaling (6)	0.83	0.83	0.83	0.83	1	0.83	0.83	0.83	1	0.83
CAR (8)	1	1	1	1	1	1	1	1	1	1
LR (5)	1	1	1	1	1	1	1	1	1	1
GLGT (3)	1	0	1	1	1	1	0	0	0	1
SRC (2)	0	1	1	1	1	1	1	1	1	1
Shearing (6)	0	0.83	0	0.17	0.33	0	0.5	0.83	0.83	0
Cropping (1)	0	1	0	1	0	0	0	0	0	0
Random bending (1)	0	0	0	1	0	0	0	0	0	0
Copy attack (2) [15]	1	1	1	1	1	1	1	1	1	1

MF: median filtering with various window sizes: 2×2, 3×3, and 4×4

GF: Gaussian filtering with window size 3×3

FMLR: Frequency Mode Laplacian Removal

JPEG: compression with quality factors, $90\% \sim 10\%$

Scaling: with factors $0.5 \sim 2.0$

CAR: change of the aspect ratio

LR: line removing (a maximum of 5 rows/columns and 17 columns/rows are removed)

GLGT: general linear geometric transformation

SRC: slight rotation ($\pm 0.25°$) and cropping

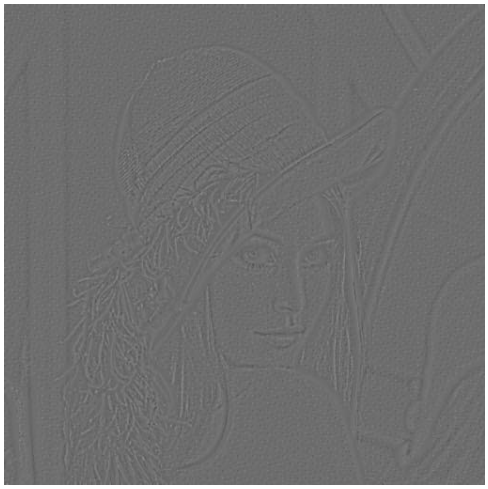
Cropping: with 1% sizes of an image discarded

Copy attack: added marks estimated by Wiener filtering with two sets of parameters

Table 3: **Verification of robustness w.r.t. combined attacks. The combined attacks were generated as JPEG2000 compression (with bit rate 1 bit/pixel, denoted as J2K) [11] + StirMark 3.1 attacks [29]. Each attack's name is followed by a digit, which indicates the number of times that the attack was performed with different parameters.**

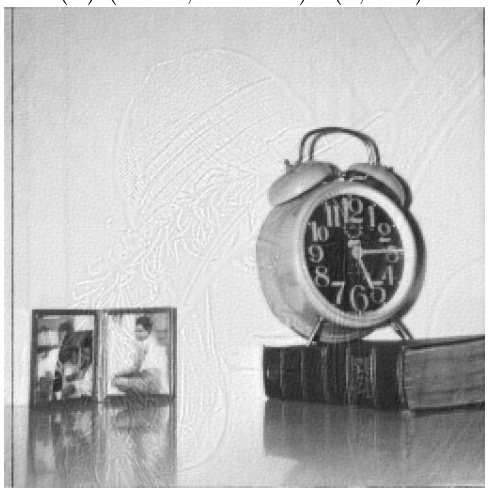
Combined attacks	3(a)	3(b)	3(c)	3(d)	3(e)	3(f)	3(g)	3(h)	3(i)	3(j)
J2K+MF (3)	1	1	1	0.67	0.33	0.67	0.67	1	1	1
J2K+GF (1)	1	1	1	1	1	1	1	1	1	1
J2K+FMLR (1)	1	1	0	1	0	1	1	1	1	1
J2K+Sharpening (1)	1	1	1	1	1	1	1	1	1	1
J2K+Color reduction (1)	1	1	1	1	0	1	1	1	1	1
J2K+JPEG (12)	0.83	0.42	0.58	0.67	0.42	0.92	0.92	1	0.92	0.83
J2K+Flipping (1)	1	1	1	1	1	0	1	1	1	1
J2K+Scaling (6)	0.83	0.83	0.83	0.83	0.67	0.83	0.83	0.83	1	0.83
J2K+CAR (8)	1	1	1	1	1	1	1	1	1	1
J2K+LR (5)	1	1	1	1	0.8	1	1	1	1	1
J2K+GLGT (3)	1	0	1	1	0	1	0	0	0	1
J2K+SRC (2)	0.5	1	0.5	0.5	0	1	0	1	1	1

(a) (mean,variance)=$(7, 90)$



(b) (mean,variance)=$(7, 180)$

(c)



(d)

(e)

(f)

Figure 5: Copy Attack: (a) and (b) show the estimated Lenna watermark; (c)∼(f) show the counterfeit images with the estimated Lenna watermark added.